RESEARCH CENTRE

Grenoble - Rhône-Alpes

IN PARTNERSHIP WITH:

Centrum Wiskunde & Informatica, Institut national des sciences appliquées de Lyon, Université Claude Bernard (Lyon 1), Université de Rome la Sapienza

2020 ACTIVITY REPORT

Project-Team ERABLE

European Research team in Algorithms and Biology, formaL and Experimental

IN COLLABORATION WITH: Laboratoire de Biométrie et Biologie Evolutive (LBBE)

DOMAIN

Digital Health, Biology and Earth

THEME

Computational Biology

Contents

Project-Team ERABLE

Creation of the Team: 2015 January 01, updated into Project-Team: 2015 July 01

Keywords

Computer sciences and digital sciences

A3. – Data and knowledge

A3.1. – Data

A3.1.1. – Modeling, representation

A3.1.4. – Uncertain data

A3.3. – Data and knowledge analysis

A3.3.2. – Data mining

A3.3.3. – Big data analysis

A7. – Theory of computation

A8.1. – Discrete mathematics, combinatorics

A8.2. – Optimization

A8.7. – Graph theory

A8.8. – Network science

A8.9. – Performance evaluation

Other research topics and application domains

B1. – Life sciences

B1.1. – Biology

B1.1.1. – Structural biology

B1.1.2. – Molecular and cellular biology

B1.1.4. – Genetics and genomics

B1.1.6. – Evolutionnary biology

B1.1.7. – Bioinformatics

B1.1.10. – Systems and synthetic biology

B2. – Health

B2.2. – Physiology and diseases

B2.2.3. – Cancer

B2.2.4. – Infectious diseases, Virology

B2.3. – Epidemiology

1 Team members, visitors, external collaborators

Research Scientists

- Marie-France Sagot [Team leader, Inria, Senior Researcher, HDR]
- Laurent Jacob [CNRS, Researcher, from Sep 2020]
- Solon Pissis [CWI - Netherlands, Researcher]
- Blerina Sinaimeri [Inria, Researcher]
- Leen Stougie [CWI - Netherlands, Senior Researcher]
- Fabrice Vavre [CNRS, Senior Researcher, HDR]
- Alain Viari [Inria, Senior Researcher]

Faculty Members

- Hubert Charles [INSA Lyon, Professor, HDR]
- Roberto Grossi [Università di Pisa - Italy, Professor]
- Giuseppe Francesco Italiano [Libera Università Internazionale degli Studi Sociali Guido Carli - Italy, Professor]
- Vincent Lacroix [Université Claude Bernard, Lyon 1, Associate Professor]
- Alberto Marchetti-Spaccamela [Sapienza Università di Roma - Italy, Professor]
- Arnaud Mary [Université Claude Bernard, Lyon 1, Associate Professor]
- Nadia Pisanti [Università di Pisa - Italy, Associate Professor]
- Cristina Vieira [Université Claude Bernard, L, Professor, HDR]

Post-Doctoral Fellows

- Audric Cologne [CNRS]
- Scheila Mucha [Inria]

PhD Students

- Marianne Borderes [MaaT Pharma Lyon, CIFRE]
- Nicolas Homberg [Inrae]
- Carol Moraga Quinteros [Conicyt - Chili, until Oct 2020]
- Camille Sessegolo [Université Claude Bernard, L]
- Antoine Villie [CNRS, from Sep 2020]
- Yishu Wang [Université Claude Bernard, Lyon 1]
- Irene Ziska [Inria, until Nov 2020]

Administrative Assistant

- Anouchka Ronceray [Inria]

2 Overall objectives

Cells are seen as the basic structural, functional and biological units of all living systems. They represent the smallest units of life that can replicate independently, and are often referred to as the building blocks of life. Living organisms are then classified into unicellular ones – this is the case of most bacteria and archea – or multicellular – this is the case of animals and plants. Actually, multicellular organisms, such as for instance human, may be seen as composed of native (human) cells, but also of extraneous cells represented by the diverse bacteria living inside the organism. The proportion in the number of the latter in relation to the number of native cells is believed to be high: this is for example of 90% in humans. Multicellular organisms have thus been described also as "superorganisms with an internal ecosystem of diverse symbiotic microbiota and parasites" (Nicholson *et al.*, Nat Biotechnol, 22(10):1268-1274, 2004) where symbiotic means that the extraneous unicellular organisms (cells) live in a close, and in this case, long-term relation both with the multicellular organisms they inhabit and among themselves. On the other hand, bacteria sometimes group into colonies of genetically identical individuals which may acquire both the ability to adhere together and to become specialised for different tasks. An example of this is the cyanobacterium *Anabaena sphaerica* who may group to form filaments of differentiated cells, some – the heterocysts – specialised for nitrogen fixation while the others are capable of photosynthesis. Such filaments have been seen as first examples of multicellular patterning.

At its extreme, one could then see life as one collection, or a collection of collections of genetically identical or distinct self-replicating cells who interact, sometimes closely and for long periods of evolutionary time, with same or distinct functional objectives. The interaction may be at equilibrium, meaning that it is beneficial or neutral to all, or it may be unstable meaning that the interaction may be or become at some time beneficial only to some and detrimental to other cells or collections of cells. The interaction may involve living systems, or systems that have been described as being at the edge of life such as viruses, or else living systems and chemical compounds (environment). It also includes the interaction between cells within a multicellular organism, or between transposable elements and their host genome.

The application goal of ERABLE is, through the use of mathematical models and algorithms, to better understand such close and often persistent interactions, with a longer term objective of becoming able in some cases to suggest the means of controlling for or of re-establishing equilibrium in an interacting community by acting on its environment or on its players, how they play and who plays. This goal requires to identify who are the partners in a closely interacting community, who is interacting with whom, how and by which means. Any model is a simplification of reality, but once selected, the algorithms to explore such model should address questions that are precisely defined and, whenever possible, be exact in the answer as well as exhaustive when more than one exists in order to guarantee an accurate interpretation of the results within the given model. This fits well the mathematical and computational expertise of the team, and drives the methodological goal of ERABLE which is to substantially and systematically contribute to the field of exact enumeration algorithms for problems that most often will be hard in terms of their complexity, and as such to also contribute to the field of combinatorics in as much as this may help in enlarging the scope of application of exact methods.

The key objective is, by constantly crossing ideas from different models and types of approaches, to look for and to infer "patterns", as simple and general as possible, either at the level of the biological application or in terms of methodology. This objective drives which biological systems are considered, and also which models and in which order, going from simple discrete ones first on to more complex continuous models later if necessary and possible.

3 Research program

3.1 Two main goals

ERABLE has two main goals, one related to biology and the other to methodology (algorithms, combinatorics, statistics). In relation to biology, the main goal of ERABLE is to contribute, through the use of mathematical models and algorithms, to a better understanding of close and often persistent interactions between "collections of genetically identical or distinct self-replicating cells" which will correspond to organisms/species or to actual cells. The first will cover the case of what has been called symbiosis, meaning when the interaction involves different species, while the second will cover the case of different collections of cells. One question in particular, but not exclusively, interests us. This is the one of a (cancerous) tumour which may be seen as a collection of cells which suddenly disrupts its interaction with the other (collections of) cells in an organism by starting to grow uncontrollably.

Such interactions are being explored initially at the molecular level. Although we rely as much as possible on already available data, we intend to also continue contributing to the identification and analysis of the main genomic and systemic (regulatory, metabolic, signalling) elements involved or impacted by an interaction, and how they are impacted. We started going to the population and ecological levels by modelling and analysing the way such interactions influence, and are or can be influenced by the ecosystem of which the "collections of cells" are a part. The key steps are:

- identifying the molecular elements based on so-called omics data (genomics, transcriptomics, metabolomics, proteomics, etc.): such elements may be gene or proteins, genetic variations, (DNA, RNA, protein) binding sites, (small and long non coding) RNAs, etc.

- simultaneously inferring and analysing the network that models how these molecular elements are physically and functionally linked together for a given goal, or find themselves associated in a response to some change in the environment;

- modelling and analysing the population and ecological network formed by the "collections of cells in interaction", meaning modelling a network of networks (previously inferred or as already available in the literature).

One important longer term goal of the above is to analyse how the behaviour and dynamics of such a network of networks might be controlled by modifying it, including by subtracting some of its components from the network or by adding new ones.

In relation to methodology, the main goal is to provide those enabling to address our main biological objective as stated above that lead to the best possible interpretation of the results within a given pre-established model and a well defined question. Ideally, given such a model and question, the method is exact and also exhaustive if more than one answer is possible. Three aspects are thus involved here: establishing the model within which questions can and will be put; clearly defining such questions; exactly answering to them or providing some guarantee on the proximity of the answer given to the "correct" one. We intend to continue contributing to these three aspects:

- at the modelling level, by exploring better models that at a same time are richer in terms of the information they contain (as an example, in the case of metabolism, using hypergraphs as models for it instead of graphs) and are susceptible to an easier treatment:

  - these two objectives (rich models that are at the same time easy to treat) might in many cases be contradictory and our intention is then to contribute to a fuller characterisation of the frontiers between the two;
  - even when feasible, the richer models may lack a full formal characterisation (this is for instance the case of hypergraphs) and our intention is then to contribute to such a characterisation;

- at the question level, by providing clear formalisations of those that will be raised by our biological concerns;

- at the answer level:

  - to extend the area of application of exact algorithms by: (i) a better exploration of the combinatorial properties of the models, (ii) the development of more efficient data structures, (iii) a smarter traversal of the space of solutions when more than one solution exists;
  - when exact algorithms are not possible, or when there is uncertainty in the input data to an algorithm, to improve the quality of the results given by a deeper exploration of the links between different algorithmic approaches: combinatorial, randomised, stochastic. More recently, the team has also become interested in exploring the interface between exact algorithms on one hand, and probabilistic or statistical ones on the other such as used in machine

learning approaches. More in particular, the team has also become interested in investigating an area of research called “interpretable machine learning” that has been developing more recently and its potential relations with exact, combinatorial approaches.

3.2 Different research axes

The goals of the team are biological and methodological, the two being intrinsically linked. Any division into axes along one or the other aspect or a combination of both is thus somewhat artificial. Following the evaluation of the team at the end of 2017, four main axes were identified, with the last one being the more recently added one. This axis is specifically oriented towards health in general. The first three axes are: (pan)genomics and transcriptomics in general, metabolism and (post)transcriptional regulation, and (co)evolution.

Notice that the division itself is based on the biological level (genomic, metabolic/regulatory, evolutionary) or main current Life Science purpose (health) rather than on the mathematical or computational methodology involved. Any choice has its part of arbitrariness. Through the one we made, we wished to emphasise the fact that the area of application of ERABLE is important for us. *It does not mean that the mathematical and computational objectives are not equally important*, but only that those are, most often, motivated by problems coming from or associated to the general Life Science goal. Notice that such arbitrariness also means that some Life Science topics will be artificially split into two different Axes. One example of this is genomics and the main health areas currently addressed that are intrinsically inter-related.

Axis 1: (Pan)Genomics and transcriptomics in general

Intra and inter-cellular interactions involve molecular elements whose identification is crucial to understand what governs, and also what might enable to control such interactions. For the sake of clarity, the elements may be classified in two main classes, one corresponding to the elements that allow the interactions to happen by moving around or across the cells, and another that are the genomic regions where contact is established. Examples of the first are non coding RNAs, proteins, and mobile genetic elements such as (DNA) transposons, retro-transposons, insertion sequences, etc. Examples of the second are DNA/RNA/protein binding sites and targets. Furthermore, both types (effectors and targets) are subject to variation across individuals of a population, or even within a single (diploid) individual. Identification of these variations is yet another topic that we wish to cover. Variations are understood in the broad sense and cover single nucleotide polymorphisms (SNPs), copy-number variants (CNVs), repeats other than mobile elements, genomic rearrangements (deletions, duplications, insertions, inversions, translocations) and alternative splicings (ASs). All three classes of identification problems (effectors, targets, variations) may be put under the general umbrella of genomic functional annotation.

Axis 2: Metabolism and (post)transcriptional regulation

As increasingly more data about the interaction of molecular elements (among which those described above) becomes available, these should then be modelled in a subsequent step in the form of networks. This raises two main classes of problems. The first is to accurately infer such networks. Assuming such a network, integrated or “simple”, has been inferred for a given organism or set of organisms, the second problem is then to develop the appropriate mathematical models and methods to extract further biological information from such networks.

The team has so far concentrated its efforts on two main aspects concerning such interactions: metabolism and post-transcriptional regulation by small RNAs. The more special niche we have been exploring in relation to metabolism concerns the fact that the latter may be seen as an organism’s immediate window into its environment. Finely understanding how species communicate through those windows, or what impact they may have on each other through them is thus important when the ultimate goal is to be able to model communities of organisms, for understanding them and possibly, on a longer term, for control. While such communication has been explored in a number of papers, most do so at a too high level or only considered couples of interacting organisms, not larger communities. The idea of investigating consortia, and in the case of synthetic biology, of using them, has thus started being developed in the last decade only, and was motivated by the fact that such consortia may perform more complicated functions than could single populations, as well as be more robust to environmental fluctuations. Another originality of the work that the team has been doing in the last decade has also been to fully explore the combinatorial aspects of the structures used (graphs or directed hypergraphs) and of the associated algorithms. As concerns post-transcriptional regulation, the team has essentially been exploring the idea that small RNAs may have an important role in the dialog between different species.

Axis 3: (Co)Evolution

Understanding how species that live in a close relationship with others may (co)evolve requires understanding for how long symbiotic relationships are maintained or how they change through time. This may have deep implications in some cases also for understanding how to control such relationships, which may be a way of controlling the impact of symbionts on the host, or the impact of the host on the symbionts and on the environment (by acting on its symbiotic partner(s)). These relationships, also called *symbiotic associations*, have however not yet been very widely studied, at least not at a large scale.

One of the problems is getting the data, meaning the trees for hosts and symbionts but even prior to that, determining with which symbionts the present-day hosts are associated (or are "infected" by as may be the term used in some contexts) which is a big enterprise in itself. The other problem is measuring the stability of the association. This has generally been done by concomitantly studying the phylogenies of hosts and symbionts, that is by doing what is called a *cophylogeny* analysis, which itself is often realised by performing what is called a *reconciliation* of two phylogenetic trees (in theory, it could be more than two but this is a problem that has not yet been addressed by the team), one for the symbionts and one for the hosts with which the symbionts are associated. This consists in mapping one of the trees (usually, the symbiont tree) to the other. Cophylogeny inherits all the difficulties of phylogeny, among which the fact that it is not possible to check the result against the "truth" as this is now lost in the past. Cophylogeny however also brings new problems of its own which are to estimate the frequency of the different types of events that could lead to discrepant evolutionary histories, and to estimate the duration of the associations such events may create.

Axis 4: Health in general

As indicated above, this is a recent axis in the team and concerns various applications to human and animal health. In some ways, it overlaps with the three previous axes as well as with Axis 5 on the methodological aspects, but since it gained more importance in the past few years, we decided to develop more these particular applications. Most of them started through collaborations with clinicians. Such applications are currently focused on three different topics: (i) Infectiology, (ii) Rare diseases, and (iii) Cancer.

Infectiology is the oldest one. It started by a collaboration with Arnaldo Zaha from the Federal University of Rio Grande do Sul in Brazil that focused on pathogenic bacteria living inside the respiratory tract of swines. Since our participation in the H2020 ITN MicroWine, we started interested in infections affecting plants this time, and more particularly vine plants. Rare Diseases on the other hand started by a collaboration with clinicians from the Centre de Recherche en Neurosciences of Lyon (CNRL) and is focused on the Taybi-Linder Syndrome (TALS) and on abnormal splicing of U12 introns, while Cancer rests on a collaboration with the Centre Léon Bérard (CLB) and Centre de Recherche en Cancérologie of Lyon (CRCL) which is focused on Breast and Prostate carcinomas and Gynaecological carcinosarcomas.

The latter collaboration was initiated through a relationship between a member of ERABLE (Alain Viari) and Dr. Gilles Thomas who had been friends since many years. G. Thomas was one of the pioneers of Cancer Genomics in France. After his death in 2014, Alain Viari took the (part time) responsibility of his team at CLB and pursued the main projects he had started.

Within Inria and beyond, the first two applications (Infectiology and Rare Diseases) may be seen as unique because of their specific focus (resp. microbiome and respiratory tract of swines / vine plants on one hand, and TALS on the other). In the first case, such uniqueness is also related to the fact that the work done involves a strong computational part but also experiments that in some cases (respiratory tract of swines) is *performed within ERABLE itself.*

4 Application domains

4.1 Biology and Health

The main areas of application of ERABLE are: (1) biology understood in its more general sense, with a special focus on symbiosis and on intracellular interactions, and (2) health with a special emphasis for now on infectious diseases, rare diseases, and cancer.

5 Social and environmental responsibility

5.1 Footprint of research activities

There are three axes on which we would like to focus in the coming years.

Travelling is essential for the team, that is European and has many international collaborations. We would however like to continue to develop as much as possible travelling by train or even car. This is something we do already, for instance between Lyon and Amsterdam by train, and that we have done in the past, such as for instance between Lyon and Pisa by car, and between Rome and Lyon by train, or even in the latter case once between Rome and Amsterdam!

Computing is also essential for the team. We would like to continue our effort to produce resource frugal software and develop better guidelines for the end users of our software so that they know better under which conditions our software is expected to be adapted, and which more resource-frugal alternatives exist, if any.

Having an impact on how data are produced is also an interest of the team. Much of the data produced is currently only superficially analysed. Generating smaller datasets and promoting data reuse could avoid not only data waste, but also economise on computer time and energy required to produce such data.

5.2 Impact of research results

As indicated earlier, the overall objective of the team is to arrive at a better understanding of close and often persistent interactions among living systems, between such living systems and viruses, between living systems and chemical compounds (environment), among cells within a multicellular organism, and between transposable elements and their host genome. There is another longer-term objective, much more difficult and riskier, a "dream" objective whose underlying motivation may be seen as social and is also environmental.

The main idea we thus wish to explore is inspired by the one universal concept underlying life. This is the concept of survival. Any living organism has indeed one single objective: to remain alive and reproduce. Not only that, any living organism is driven by the need to give its descendants the chance to perpetuate themselves. As such, no organism, and more in general, no species can be considered as "good" or "bad" in itself. Such concepts arise only from the fact that resources, some of which may be shared among different species, are of limited availability. Conflict thus seems inevitable, and "war" among species the only way towards survival.

However, this is not true in all cases. Conflict is often observed, even actively pursued by, for instance, humans. Two striking examples that have been attracting attention lately, not necessarily in a way that is positive for us, are related to the use of antibiotics on one hand, and insecticides on the other, both of which, especially but not only the second can also have disastrous environmental consequences. Yet cooperation, or at least the need to stop distinguishing between "good" (mutualistic) and "bad" (parasitic) interactions appears to be, and indeed in many circumstances is of crucial importance for survival. The two questions which we want to address are: (i) what happens to the organisms involved in "bad" interactions with others (for instance, their human hosts) when the current treatments are used, and (ii) can we find a non-violent or cooperative way to treat such diseases?

Put in this way, the question is infinitely vast. It is not completely utopic. We had the opportunity in recent years to discuss such question with notably biologists with whom we were involved in two European projects (namely BachBerry, http://team.inria.fr/erable/en/older-projects/fp7-kbbe-bachberry/, and MicroWine, http://team.inria.fr/erable/en/older-projects/microwi ne/). In both cases, we had examples of bacteria that are "bad" when present in a certain environment, and "good" when the environment changes. In one of the cases at least, related to vine plants, such change in environment seems to be related to the presence of other bacteria. This idea is already explored in agriculture to avoid the use of insecticide. Such exploration is however still relatively limited in terms of scope, and especially, has not yet been fully investigated scientifically.

The aim will be to reach some proofs of concepts, which may then inspire others, including ourselves on a longer term, to pursue research along this line of thought. Such proofs will in themselves already require to better understand what is involved in, and what drives or influences any interaction.

6 Highlights of the year

The research of all team members, in particular of PhD students or Postdocs, is important for us and we prefer not to highlight any in particular.

7 New software and platforms

We indicate in this section the new methods we developed in 2020 but also the older ones that continue to be used and that are being constantly maintained by the researchers of the team. This indeed represents a great part of our effort and time, and is important in general.

7.1 New software

7.1.1 BrumiR

Name: A toolkit for de novo discovery of microRNAs from sRNA-seq data.

Keywords: Bioinformatics, Structural Biology, Genomics

Functional Description: BRUMIR is an algorithm that is able to discover miRNAs directly and exclusively from sRNA-seq data. It was benchmarked with datasets encompassing animal and plant species using real and simulated sRNA-seq experiments. The results show that BRUMIR reaches the highest recall for miRNA discovery, while at the same time being much faster and more efficient than the state-of-the-art tools evaluated. The latter allows BRUMIR to analyse a large number of sRNA-seq experiments, from plant or animal species. Moreover, BRUMIR detects additional information regarding other expressed sequences (sRNAs, isomiRs, etc.), thus maximising the biological insight gained from sRNA-seq experiments. Finally, when a reference genome is available, BRUMIR provides a new mapping tool (BRUMIR2REFERENCE) that performs a posteriori an exhaustive search to identify the precursor sequences.

URL: https://github.com/camoragaq/BrumiR

Contacts: Carol Moraga Quinteros, Marie-France Sagot

Participants: Carol Moraga Quinteros, Marie-France Sagot

7.1.2 Capybara

Name: equivalence ClAss enumeration of coPhylogenY event-BAsed ReconciliAtions

Keywords: Bioinformatics, Evolution

Functional Description: Phylogenetic tree reconciliation is the method of choice in analysing host-symbiont systems. Despite the many reconciliation tools that have been proposed in the literature, two main issues remain unresolved: listing suboptimal solutions (*i.e.*, whose score is “close” to the optimal ones), and listing only solutions that are biologically different “enough”. The first issue arises because the optimal solutions are not always the ones biologically most significant, providing many suboptimal solutions as alternatives for the optimal ones is thus very useful. The second one is related to the difficulty to analyse an often huge number of optimal solutions. Capybara addresses both of these problems in an efficient way. Furthermore, it includes a tool for visualising the solutions that significantly helps the user in the process of analysing the results.

URL: `https://github.com/Helio-Wang/Capybara-app`

Publication: hal-02917341

Contacts: Yishu Wang, Arnaud Mary, Marie-France Sagot, Blerina Sinaimeri

Participants: Yishu Wang, Arnaud Mary, Marie-France Sagot, Blerina Sinaimeri

7.1.3 C3Part/Isofun

Keywords: Bioinformatics, Genomics

Functional Description: The C3PART / ISOFUN package implements a generic approach to the local alignment of two or more graphs representing biological data, such as genomes, metabolic pathways or protein-protein interactions, in order to infer a functional coupling between them.

URL: `http://www.inrialpes.fr/helix/people/viari/lxgraph/index.html`

Contact: Alain Viari

Participants: Alain Viari, Anne Morgat, Frédéric Boyer, Marie-France Sagot, Yves-Pol Deniélou

7.1.4 Cassis

Keywords: Bioinformatics, Genomics

Functional Description: Implements methods for the precise detection of genomic rearrangement breakpoints.

URL: `http://pbil.univ-lyon1.fr/software/Cassis/`

Contact: Marie-France Sagot

Participants: Christian Baudet, Christian Gautier, Claire Lemaitre, Eric Tannier, Marie-France Sagot

7.1.5 Coala

Name: CO-evolution Assessment by a Likelihood-free Approach

Keywords: Bioinformatics, Evolution

Functional Description: COALA stands for “COevolution Assessment by a Likelihood-free Approach”. It is thus a likelihood-free method for the co-phylogeny reconstruction problem which is based on an Approximate Bayesian Computation (ABC) approach.

URL: `http://team.inria.fr/erable/en/software/coala/`

Contact: Blerina Sinaimeri

Participants: Beatrice Donati, Blerina Sinaimeri, Catherine Matias, Christian Baudet, Christian Gautier, Marie-France Sagot, Pierluigi Crescenzi

7.1.6 CSC

Keywords: Genomics, Algorithm

Functional Description: Given two sequences x and y, CSC (which stands for Circular Sequence Comparison) finds the cyclic rotation of x (or an approximation of it) that minimises the blockwise q-gram distance from y.

URL: https://github.com/solonas13/csc

Contact: Nadia Pisanti

7.1.7 Cycads

Keywords: Systems Biology, Bioinformatics

Functional Description: Annotation database system to ease the development and update of enriched BIOCYC databases. CYCADS allows the integration of the latest sequence information and functional annotation data from various methods into a metabolic network reconstruction. Functionalities will be added in future to automate a bridge to metabolic network analysis tools, such as METEXPLORE. CYCADS was used to produce a collection of more than 22 arthropod metabolism databases, available at ACYPICYC (http://acypicyc.cycadsys.org) and ARTHROPODACYC (http://arthropodacyc.cycadsys.org). It will continue to be used to create other databases (newly sequenced organisms, Aphid biotypes and symbionts...).

URL: http://www.cycadsys.org/

Contact: Hubert Charles

Participants: Augusto Vellozo, Hubert Charles, Marie-France Sagot, Stefano Colella

7.1.8 DBGWAS

Keywords: Graph algorithmics, Genomics

Functional Description: DBGWAS is a tool for quick and efficient bacterial GWAS. It uses a compacted De Bruijn Graph (cDBG) structure to represent the variability within all bacterial genome assemblies given as input. Then cDBG nodes are tested for association with a phenotype of interest and the resulting associated nodes are then re-mapped on the cDBG. The output of DBGWAS consists of regions of the cDBG around statistically significant nodes with several informations related to the phenotypes, offering a representation helping in the interpretation. The output can be viewed with any modern web browser, and thus easily shared.

URL: https://gitlab.com/leoisl/dbgwas

Contact: Laurent Jacob

7.1.9 Eucalypt

Keywords: Bioinformatics, Evolution

Functional Description: EUCALYPT stands for "EnUmerator of Coevolutionary Associations in PoLYnomial-Time delay". It is an algorithm for enumerating all optimal (possibly time-unfeasible) mappings of a symbiont tree unto a host tree.

URL: http://team.inria.fr/erable/en/software/eucalypt/

Contact: Blerina Sinaimeri

Participants: Beatrice Donati, Blerina Sinaimeri, Christian Baudet, Marie-France Sagot, Pierluigi Crescenzi

7.1.10 Fast-SG

Keywords: Genomics, Algorithm, NGS

Functional Description: FAST-SG enables the optimal hybrid assembly of large genomes by combining short and long read technologies.

URL: https://github.com/adigenova/fast-sg

Contacts: Alex Di Genova, Marie-France Sagot

Participants: Alex Di Genova, Marie-France Sagot, Alejandro Maass, Gonzalo Ruz Heredia

7.1.11 Gobbolino-Touché

Keywords: Bioinformatics, Graph algorithmics, Systems Biology

Functional Description: Designed to solve the metabolic stories problem, which consists in finding all maximal directed acyclic subgraphs of a directed graph G whose sources and targets belong to a subset of the nodes of G, called the black nodes.

URL: http://gforge.inria.fr/projects/gobbolino

Contact: Marie-France Sagot

Participants: Etienne Birmelé, Fabien Jourdan, Ludovic Cottret, Marie-France Sagot, Paulo Vieira Milreu, Pierluigi Crescenzi, Vicente Acuna Aguayo, Vincent Lacroix

7.1.12 HapCol

Keywords: Bioinformatics, Genomics

Functional Description: A fast and memory-efficient DP approach for haplotype assembly from long reads that works until 25x coverage and solves a constrained minimum error correction problem exactly.

URL: http://hapcol.algolab.eu/

Contact: Nadia Pisanti

7.1.13 HgLib

Name: HyperGraph Library

Keywords: Graph algorithmics, Hypergraphs

Functional Description: The open-source library hglib is dedicated to model hypergraphs, which are a generalisation of graphs. In an *undirected* hypergraph, an hyperedge contains any number of vertices. A *directed* hypergraph has hyperarcs which connect several tail and head vertices. This library, which is written in C++, allows to associate user defined properties to vertices, to hyperedges/hyperarcs and to the hypergraph itself. It can thus be used for a wide range of problems arising in operations research, computer science, and computational biology.

Release Contributions: Initial version

URL: https://gitlab.inria.fr/kirikomics/hglib

Authors: Martin Wannagat, David P. Parsons, Arnaud Mary

Contacts: Arnaud Mary, David P. Parsons

Participants: Martin Wannagat, David P. Parsons, Arnaud Mary, Irene Ziska

7.1.14 KissDE

Keywords: Bioinformatics, NGS

Functional Description: KISSDE is an R Package enabling to test if a variant (genomic variant or splice variant) is enriched in a condition. It takes as input a table of read counts obtained from an NGS data pre-processing and gives as output a list of condition-specific variants.

Release Contributions: This new version improved the recall and made more precise the size of the effect computation.

URL: http://kissplice.prabi.fr/tools/kissDE/

Contact: Vincent Lacroix

Participants: Camille Marchet, Aurélie Siberchicot, Audric Cologne, Clara Benoît-Pilven, Janice Kielbassa, Lilia Brinza, Vincent Lacroix

7.1.15 KisSplice

Keywords: Bioinformatics, Bioinfirmatics search sequence, Genomics, NGS

Functional Description: Enables to analyse RNA-seq data with or without a reference genome. It is an exact local transcriptome assembler, which can identify SNPs, indels and alternative splicing events. It can deal with an arbitrary number of biological conditions, and will quantify each variant in each condition.

Release Contributions: Improvements : The KissReads module has been modified and sped up, with a significant impact on run times. Parameters : –timeout default now at 10000: in big datasets, recall can be increased while run time is a bit longer. Bugs fixed : –Reads containing only 'N': the graph construction was stopped if the file contained a read composed only of 'N's. This is was a silence bug, no error message was produced. –Problems compiling with new versions of MAC OSX (10.8+): KisSplice is now compiling with the new default C++ compiler of OSX 10.8+.

KISSPLICE was applied to a new application field, virology, through a collaboration with the group of Nadia Naffakh at Institut Pasteur. The goal is to understand how a virus (in this case influenza) manipulates the splicing of its host. This led to new developments in KISSPLICE. Taking into account the strandedness of the reads was required, in order not to mis-interpret transcriptional readthrough. We now use BCALM instead of DBG-V4 for the de Bruijn graph construction and this led to major improvements in memory and time requirements of the pipeline. We still cannot scale to very large datasets like in cancer, the time limiting step being the quantification of bubbles.

URL: http://kissplice.prabi.fr/

Authors: Gustavo Akio Tominaga Sacomoto, Vincent Lacroix, Pierre Peterlongo, Rayan Chikhi, Alice Julien-Laferrière, David P. Parsons, Janice Kielbassa, Marie-France Sagot, Pavlos Antoniou, Uricaru Raluca, Vincent Miele

Contact: Vincent Lacroix

Participants: Alice Julien-Laferrière, Leandro Ishi Soares de Lima, Vincent Miele, Rayan Chikhi, Pierre Peterlongo, Camille Marchet, Gustavo Akio Tominaga Sacomoto, Marie-France Sagot, Vincent Lacroix

7.1.16 KisSplice2RefGenome

Keywords: Bioinformatics, NGS, Transcriptomics

Functional Description: KISSPLICE identifies variations in RNA-seq data, without a reference genome. In many applications however, a reference genome is available. KISSPLICE2REFGENOME enables to facilitate the interpretation of the results of KISSPLICE after mapping them to a reference genome.

URL: http://kissplice.prabi.fr/tools/kiss2refgenome/

Contact: Vincent Lacroix

Participants: Audric Cologne, Camille Marchet, Camille Sessegolo, Alice Julien-Laferrière, Vincent Lacroix

7.1.17 KisSplice2RefTranscriptome

Keywords: Bioinformatics, NGS, Transcriptomics

Functional Description: KisSplice2RefTranscriptome enables to combine the output of KisSplice with the output of a full length transcriptome assembler, thus allowing to predict a functional impact for the positioned SNPs, and to intersect these results with condition-specific SNPs. Overall, starting from RNA-seq data only, we obtain a list of condition-specific SNPs stratified by functional impact.

URL: http://kissplice.prabi.fr/tools/kiss2rt/

Contact: Vincent Lacroix

Participants: Helene Lopez Maestre, Mathilde Boutigny, Vincent Lacroix

7.1.18 MetExplore

Keywords: Systems Biology, Bioinformatics

Functional Description: Web-server that allows to build, curate and analyse genome-scale metabolic networks. MetExplore is also able to deal with data from metabolomics experiments by mapping a list of masses or identifiers onto filtered metabolic networks. Finally, it proposes several functions to perform Flux Balance Analysis (FBA). The web-server is mature, it was developed in PHP, JAVA, Javascript and Mysql. MetExplore was started under another name during Ludovic Cottret's PhD in Bamboo, and is now maintained by the MetExplore group at the Inra of Toulouse.

URL: https://metexplore.toulouse.inra.fr/index.html/

Contacts: Fabien Jourdan, Ludovic Cottret, Marie-France Sagot

Participants: Fabien Jourdan, Hubert Charles, Ludovic Cottret, Marie-France Sagot

7.1.19 Mirinho

Keywords: Bioinformatics, Computational biology, Genomics, Structural Biology

Functional Description: Predicts, at a genome-wide scale, microRNA candidates.

URL: http://team.inria.fr/erable/en/software/mirinho/

Contact: Marie-France Sagot

Participants: Christian Gautier, Christine Gaspin, Cyril Fournier, Marie-France Sagot, Susan Higashi

7.1.20 Momo

Name: Multi-Objective Metabolic mixed integer Optimization

Keywords: Metabolism, Metabolic networks, Multi-objective optimisation

Functional Description: Momo is a multi-objective mixed integer optimisation approach for enumerating knockout reactions leading to the overproduction and/or inhibition of specific compounds in a metabolic network.

URL: http://team.inria.fr/erable/en/software/momo/

Contacts: Marie-France Sagot, Susana Vinga, Ricardo Luiz de Andrade Abrantes

Participants: Ricardo Luiz de Andrade Abrantes, Nuno Mira, Susana Vinga, Marie-France Sagot

7.1.21 Moomin

Name: Mathematical explOration of Omics data on a MetabolIc Network

Keywords: Metabolic networks, Transcriptomics

Functional Description: MOOMIN is a tool for analysing differential expression data. It takes as its input a metabolic network and the results of a DE analysis: a posterior probability of differential expression and a (logarithm of a) fold change for a list of genes. It then forms a hypothesis of a metabolic shift, determining for each reaction its status as "increased flux", "decreased flux", or "no change". These are expressed as colours: red for an increase, blue for a decrease, and grey for no change. See the paper for full details: https://doi.org/10.1093/bioinformatics/btz584

URL: https://github.com/htpusa/moomin

Contact: Marie-France Sagot

Participants: Henri Taneli Pusa, Mariana Ferrarini, Ricardo Luiz de Andrade Abrantes, Arnaud Mary, Alberto Marchetti-Spaccamela, Leen Stougie, Marie-France Sagot

7.1.22 MultiPus

Keywords: Systems Biology, Algorithm, Graph algorithmics, Metabolic networks, Computational biology

Functional Description: MULTIPUS (for “MULTIple species for the synthetic Production of Useful biochemical Substances”) is an algorithm that, given a microbial consortium as input, identifies all optimal sub-consortia to synthetically produce compounds that are either exogenous to it, or are endogenous but where interaction among the species in the sub-consortia could improve the production line.

URL: https://team.inria.fr/erable/en/software/multipus/

Contacts: Marie-France Sagot, Arnaud Mary

Participants: Alberto Marchetti-Spaccamela, Alice Julien-Laferrière, Arnaud Mary, Delphine Parrot, Laurent Bulteau, Leen Stougie, Marie-France Sagot, Susana Vinga

7.1.23 Pitufolandia

Keywords: Bioinformatics, Graph algorithmics, Systems Biology

Functional Description: The algorithms in PITUFOLANDIA (PITUFO / PITUFINA / PAPAPITUFO) are designed to solve the minimal precursor set problem, which consists in finding all minimal sets of precursors (usually, nutrients) in a metabolic network that are able to produce a set of target metabolites.

URL: http://gforge.inria.fr/projects/pitufo/

Contact: Marie-France Sagot

Participants: Vicente Acuna Aguayo, Paulo Vieira Milreu, Alberto Marchetti-Spaccamela, Leen Stougie, Martin Wannagat, Marie-France Sagot

7.1.24 Sasita

Keywords: Bioinformatics, Graph algorithmics, Systems Biology

Functional Description: SASITA is a software for the exhaustive enumeration of minimal precursor sets in metabolic networks.

URL: https://team.inria.fr/erable/en/software/sasita/

Contact: Marie-France Sagot

Participants: Vicente Acuna Aguayo, Ricardo Luiz de Andrade Abrantes, Paulo Vieira Milreu, Alberto Marchetti-Spaccamela, Leen Stougie, Martin Wannagat, Marie-France Sagot

7.1.25 Smile

Keywords: Bioinformatics, Genomic sequence

Functional Description: Motif inference algorithm taking as input a set of biological sequences.

Contact: Marie-France Sagot

Participant: Marie-France Sagot

7.1.26 Rime

Keywords: Bioinformatics, Genomics, Sequence alignment

Functional Description: Detects long similar fragments occurring at least twice in a set of biological sequences.

Contacts: Nadia Pisanti, Marie-France Sagot

Participants: Nadia Pisanti, Marie-France Sagot

7.1.27 Totoro

Name: Transient respOnse to meTabOlic pertuRbation inferred at the whole netwOrk level

Keywords: Bioinformatics, Graph algorithmics, Systems Biology

Functional Description: TOTORO is a constraint-based approach that integrates internal metabolite concentrations that were measured before and after a perturbation into genome-scale metabolic reconstructions. It predicts reactions that were active during the transient state that occurred after the perturbation. The method is solely based on metabolomic data.

URL: https://gitlab.inria.fr/erable/totoro

Contacts: Irene Ziska, Arnaud Mary, Marie-France Sagot

Participants: Irene Ziska, Arnaud Mary, Marie-France Sagot

7.1.28 Wengan

Name: Making the path

Keyword: Genome assembly

Functional Description: WENGAN is a new genome assembler that unlike most of the current long-reads assemblers avoids entirely the all-vs-all read comparison. The key idea behind WENGAN is that long-read alignments can be inferred by building paths on a sequence graph. To achieve this, WENGAN builds a new sequence graph called the Synthetic Scaffolding Graph. The SSG is built from a spectrum of synthetic mate-pair libraries extracted from raw long-reads. Longer alignments are then built by performing a transitive reduction of the edges. Another distinct feature of WENGAN is that it performs self-validation by following the read information. WENGAN identifies miss-assemblies at differents steps of the assembly process.

URL: `https://github.com/adigenova/wengan`

Contacts: Marie-France Sagot, Alex Di Genova

Participants: Alex Di Genova, Marie-France Sagot

7.1.29 WhatsHap

Keywords: Bioinformatics, Genomics

Functional Description: WHATSHAP is a DP approach for haplotype assembly from long reads that works until 20x coverage and solves the minimum error correction problem exactly. PWHATSHAP is a parallelisation of the core dynamic programming algorithm of WHATSHAP.

URL: `https://bitbucket.org/whatshap/whatshap`

Contact: Nadia Pisanti

7.2 New platforms

No platforms for now due to a lack of human means to develop and especially to maintain one. An initial one is however actively planned for the coming years.

8 New results

8.1 General comments

We present in this section the main results obtained in 2020.

We tried to organise these along the four axes as presented above. Clearly, in some cases, a result obtained overlaps more than one axis. In such case, we chose the one that could be seen as the main one concerned by such results.

On the other hand, we chose not to detail the results on more theoretical aspects of computer science when these are initially addressed in contexts not directly related to computational biology even though those on string [29, 1, 4, 36, 33, 20] and graph algorithms in general [5, 14, 34] are relevant for life sciences, such as for instance (pan)genome analysis, or could become more specifically so in a near future.

A few other results of 2020 are not mentioned in this report, not because the corresponding work is not important, but because it was likewise more specialised [35]. In the same way, also for space reasons, we chose not to detail the results presented in some biological papers of the team when these did not require a mathematical or algorithmic input [6, 11, 12, 13, 18, 19, 21].

On the other hand, we do mention a work that was in revision and indeed accepted just at the end of 2020 but ended up being published at the very beginning of 2021.

Finally, we wish to call attention to the fact that some members of ERABLE, at CWI and at the University of Pisa, have been working on a theoretical problem which is important in relation to our main area of application. This problem indeed concerns privacy of the information that may be inferred by some of the algorithms developed, and more precisely what has been called string sanitization [8, 31, 32].

8.2 Axis 1: (Pan)Genomics and transcriptomics in general

Alternative splicing and variant detection

In a paper published in *BMC Bioinformatics* [23], we introduced a new algorithm and the corresponding tool EBWT2INDEL that extends our own previous framework (Prezza *et al.*, Algorithms for Molecular Biology, 14(1): 1-13, 2019) to detect also INDELs, and implements recent findings that allow to perform the whole analysis using just a Burrows-Wheeler Transform, thus reducing the working space by one order of magnitude and enabling the analysis of full genomes. We also describe a simple strategy for effectively parallelising our tool for SNP detection only. On a 24-cores machine, the parallel version of our tool is one order of magnitude faster than the sequential one. The tool is available at `https://github.com/nicolaprezza/ebwt2InDel`. The results obtained on a synthetic dataset covered at 30x (human chromosome 1) show that our tool is indeed able to find up to 83% of the SNPs and 72% of the existing INDELs. These percentages considerably improve the 71% of SNPs and 51% of INDELs found by the state-of-the art tools based on de Bruijn graphs. We furthermore reported results on larger (real) human whole-genome sequencing experiments. In such cases also, our tool exhibits a higher sensitivity than the state-of-the art tools.

Still on the issue of variants, we studied the biallelic variants in RNU4ATAC, a non-coding gene transcribed into the minor spliceosome component U4atac snRNA which are responsible for three rare recessive developmental diseases, namely Taybi-Linder/MOPD1, Roifman and Lowry-Wood syndromes. Next-generation sequencing of clinically heterogeneous cohorts (children with either a suspected genetic disorder or a congenital microcephaly) recently identified mutations in this gene, illustrating how profoundly these technologies are modifying genetic testing and assessment. As RNU4ATAC has a single non-coding exon, the bioinformatic prediction algorithms assessing the effect of sequence variants on splicing or protein function are irrelevant, which makes variant interpretation challenging to molecular diagnostic laboratories. In order to facilitate and improve clinical diagnostic assessment and genetic counselling, we presented i) an update of the previously reported RNU4ATAC mutations and an analysis of the genetic variations affecting this gene using the Genome Aggregation Database (gnomAD) resource; ii) the pathogenicity prediction performances of scores computed based on an RNA structure prediction tool and of those produced by the Combined Annotation Dependent Depletion tool for the 285 RNU4ATAC variants identified in patients or in large-scale sequencing projects; iii) a method, based on a cellular assay, that allows to measure the effect of RNU4ATAC variants on splicing efficiency of a minor (U12-type) reporter intron. Lastly, the concordance of the bioinformatic predictions and cellular assay results was investigated. This work was published in *PLoS One* [7].

Finally, again on the issue of alternative splicing events, we studied influenza A viruses (IAVs) which use diverse mechanisms to interfere with cellular gene expression. Although many RNA-seq studies had previously documented IAV-induced changes in host mRNA abundance, few were designed to allow an accurate quantification of changes in host mRNA splicing. Here, we showed that IAV infection of human lung cells induces widespread alterations of cellular splicing, with an overall increase in exon inclusion and decrease in intron retention. Over half of the mRNAs that show differential splicing undergo no significant changes in abundance or in their 3' end termination site, suggesting that IAVs can specifically manipulate cellular splicing. Among a randomly selected subset of 21 IAV-sensitive alternative splicing events, most are specific to IAV infection as they are not observed upon infection with VSV, induction of interferon expression or induction of an osmotic stress. Finally, the analysis of splicing changes in RED-depleted cells revealed a limited but significant overlap with the splicing changes in IAV-infected cells. This observation suggests that hijacking of RED by IAVs to promote splicing of the abundant viral NS1 mRNAs could partially divert RED from its target mRNAs. This work was published in *Nar Genomics and Bioinformatics* [3]. All our RNA-seq datasets and analyses are made accessible for browsing through a user-friendly Shiny interface (`http://virhostnet.prabi.fr:3838/shinyapps/flu-splicing` or `https://github.com/cbenoitp/flu-splicing`).

Bubble generator

Bubbles are pairs of internally vertex-disjoint (s,t)-paths in a directed graph. In de Bruijn graphs built from reads of RNA and DNA data, bubbles represent interesting biological events, such as alternative splicing (AS) and allelic differences (SNPs and indels). However, the set of all bubbles in a de Bruijn graph built from real data is usually too large to be efficiently enumerated and analysed in practice. In particular, despite significant research done in this area, listing bubbles still remains the main bottleneck for tools that detect AS events in a reference-free context. We recently introduced the concept of a bubble generator as a way for obtaining a compact representation of the bubble space of a graph (Acuña *et al.*, *Algorithmica*, 82:898-914, 2019, appeared in 2020). Although this generator was quite effective in finding AS events, preliminary experiments showed that it is about 5 times slower than state-of-art methods. This year, we proposed a new family of bubble generators which improve substantially on the previous one. Indeed, the generators in this new family are about two orders of magnitude faster and are still able to achieve similar precision in identifying AS events. To highlight the practical value of our new generators, we also reported some experimental results on a real dataset. This work was presented at IWOCA [28].

Genome assembly

Generating accurate genome assemblies of large, repeat-rich human genomes has proved difficult using only long, error-prone reads, and most human genomes assembled from long reads add accurate short reads to polish the consensus sequence. We developed a hybrid assembly method, that we called WENGAN, which provides the highest quality at a low computational cost. We applied WENGAN to perform a de novo assembly of four human genomes using a combination of sequencing data generated on ONT PromethION, PacBio Sequel, Illumina and MGI technology. WENGAN implements efficient algorithms to improve assembly contiguity as well as consensus quality. The resulting genome assemblies thus have high contiguity (contig NG50:17.24-80.64 Mb), few assembly errors (contig NGA50:11.8-59.59 Mb), good consensus quality (QV:27.84-42.88), and high gene completeness (BUSCO complete: 94.6-95.2%), while consuming low computational resources (CPU hours:187-1,200). In particular, the assembly of the haploid CHM13 sample achieved a contig NG50 of 80.64 Mb (NGA50:59.59 Mb), which surpasses the contiguity of the current human reference genome (GRCh38 contig NG50:57.88 Mb). The paper presentinh WENGAN was published in *Nature Biotechnology* [16]. WENGAN is available at `https://github.com/adigenova/wengan`.

On the same topic still, we also worked in the context of an haplotype-aware genome assembly whose aim is to reconstruct all individual haplotypes from a mixed sample and to provide the corresponding abundance estimates. We developed a reference-genome-independent solution based on the construction of a variation graph that captures all the diversity present in the sample. We solved the contig abundance estimation problem and proposed a greedy algorithm to efficiently build maximal-length haplotypes. We then obtained accurate frequency estimates for the reconstructed haplotypes using linear programming techniques. Our method outperforms the state-of-the-art approaches on viral quasispecies benchmarks and has the potential to assemble bacterial genomes in a strain-aware manner as well. This work was presented at RECOMB [30].

Binning

The human gut microbiota performs functions that are essential for the maintenance of the host physiology. However, characterising the functioning of microbial communities in relation to the host remains challenging in reference-based metagenomic analyses. Indeed, as taxonomic and functional analyses are performed independently, the link between the genes and the species remains unclear. Although a first set of species-level bins was built by clustering co-abundant genes, no reference bin set is established on the most used gut microbiota catalog, the Integrated Gene Catalog (IGC). With the aim to identify the best suitable method to group the IGC genes, we benchmarked nine taxonomy-independent binners implementing abundance-based, hybrid and integrative approaches. To this purpose, we designed a Simulated non-redundant Gene Catalog (SGC) and computed adapted assessment metrics. We showed that, overall, the best trade-off between the main metrics is reached by an integrative binner. For each approach, we then compared the results of the best-performing binner with our expected community structures and applied the method to IGC. We showed that the three approaches are distinguished by specific advantages, and by inherent or scalability limitations. We conclude from this that hybrid and integrative binners show promising and potentially complementary results but require improvements to be used on IGC to recover human gut microbial species. This work was submitted to *NAR Genomics and Bioinformatics* in 2020, and is now accepted. It will be published in early 2021. This is the work of the PhD student Marianne Borderes, co-supervised by M.-F. Sagot and S. Vinga (Instituto Superior Técnico, Lisbon), and funded by the ANR Technology 9.1 with the company MaatPharma, under the supervision of Lilia Boucinha initially, and then since 2019 of Emmanuel Prestat. This work with S. Vinga was part of the Inria Associated Team 10.1.1 which lasted from 2018 until this year (2020).

8.3 Axis 2: Metabolism and (post)transcriptional regulation

Metabolism

In a paper published in *BMC Bioinformatics* [2], we explored the concept of multi-objective optimisation in the field of metabolic engineering when both continuous and integer decision variables are involved in the model. In particular, we proposed a multi-objective model that may be used to suggest reaction deletions that maximise and/or minimise several functions simultaneously. The applications may include, among others, the concurrent maximisation of a bioproduct and of biomass, or maximisation of a bioproduct while minimising the formation of a given by-product, two common requirements in microbial metabolic engineering. Production of ethanol by the widely used cell factory *Saccharomyces cerevisiae* (Yeast) was adopted as a case study to demonstrate the usefulness of the proposed approach in identifying genetic manipulations that improve productivity and yield of this economically highly relevant bioproduct. We did an *in vivo* validation and we could show that some of the predicted deletions exhibit increased ethanol levels in comparison with the wild-type strain. The multi-objective programming framework we developed, called MOMO, is open-source and uses POLYSCIP (available at `http://polyscip.zib.de/`) as underlying multi-objective solver. MOMO itself is available at `https://team.inria.fr/erable/en/software/momo/`. This work was done with S. Vinga and N. Mira, both at the Instituto Superior Técnico of Lisbon, in Portugal, and was part of the project proposed within the Inria Associated Team 10.1.1.

This work was then continued with a PhD student, Irene Ziska, funded by Inria and co-supervised by M.-F. Sagot and S. Vinga. It was also part of the project of the Inria Associated Team 10.1.1. In the case of Irene's work, and always in collaboration also with N. Mira, we have developed a new method that identifies potential metabolic engineering strategies such as gene and reaction knock-outs as did MOMO, however this time does also explicitly take into account that in some cases, the target chemical can be toxic for the microorganism itself, which might render the production unstable. This new method thus aims to identify knock-outs which increase the production of the target and which, at the same time, ensure that the microorganism keeps a high resistance against the toxic target. In a first step, our approach uses multi-objective linear optimisation to find valid trade-offs between growth, target production and toxicity resistance against the target. Afterwards, potential knock-outs are enumerated and then ranked to choose the best candidates for a desired trade-off. The toxicity resistance is measured by the activity of a set of critical reactions that have to be known or identified experimentally as a prerequisite. To test our method, we applied it to identify knock-outs for the production of ethanol in Yeast. A paper is being prepared to be submitted in early 2021.

Finally, still on metabolism, we also submitted an article that presents a novel computational method called TOTORO (for "Transient respOnse to meTabOlic pertuRbation inferred at the whole netwOrk level") and which integrates internal metabolites concentrations that were measured before and after a perturbation into genome-scale metabolic reconstructions in order to predict the reactions that were active during the transient state that occurred after the perturbation. The proposed method is a constraint-based approach that takes the stoichiometry of the network into account. It minimises the change in concentrations for unmeasured metabolites and also the number of active reactions during transient state to account for a parsimonious assumption. An implementation in C++ is freely available at `https://gitlab.inria.fr/erable/totoro`. TOTORO is able to handle full networks and to consider in the model stoichiometry, cycles, reversible reactions as well as co-factors. This work is also part of Irene Ziska's PhD [38], and of the Inria Associated Team project 10.1.1.

Post-transcriptional regulation

MicroRNAs (miRNAs) belong to a class of small non-coding RNAs (ncRNAs) of 18-24 nucleotides in part responsible for post-transcriptional gene regulation in eukaryotes. These evolutionarily conserved molecules influence fundamental biological processes, including cell proliferation, differentiation, apoptosis, immune response, and metabolism. Accurately identifying miRNAs has however proven difficult. In the last decade, with the increasing accessibility of high-throughput sequencing technologies, different methods have been developed to identify miRNAs, but most of them rely exclusively on pre-existing reference genomes. Despite all the advancements in the sequencing technologies and de novo assembly algorithms, few complete genomes are available today. This represents a recurrent problem for researchers working on non-model species. The lack of a high-quality reference genome thus reduces the possibilities for discovering novel miRNAs. In a paper currently under revision, we introduced BRUMIR, which is a package composed of three tools; 1) a new discovery miRNA tool (BRUMIR-CORE) a specific genome mapper (BRUMIR2REFERENCE), and 3) a sRNA-seq read simulator (MIRSIM). In particular, BRUMIR-CORE is a *de novo* algorithm based on a de Bruijn graph approach that is able to identify miRNAs directly and exclusively from sRNA-seq data. Unlike other state-of-the-art tools, BRUMIR does not rely on a reference genome, on the availability of close phylogenetic species, or on conserved sequence information. Instead, BRUMIR starts from a de Bruijn graph encoding all the reads and is able to directly identify putative mature miRNAs on the generated graph. Along with miRNA discovery, BRUMIR assembles and identifies other types of small and long non-coding RNAs expressed within the sequencing data. Additionally, when a reference genome is available, BRUMIR provides a new mapping tool (BRUMIR2REFERENCE) that performs an exhaustive search to identify and validate the precursor sequences. We extensively benchmarked the BRUMIR toolkit on animal and plant species using simulated and real datasets. The benchmark results show that BRUMIR is very sensitive, is the fastest tool, and its predictions were supported by the characteristic hairpin structure of miRNAs. Finally, we showed the power of BRUMIR for discovering novel miRNAs in the model plant *Arabidopsis thaliana*. We sequenced a total of 18 sRNA-seq libraries from different stages of root development and used the BrumiR toolkit to analyze our data. We annotated three novel miRNAs involved in root development, showing on a real biological situation how BRUMIR catches novel information even in the case of highly annotated genomes. The paper presenting BRUMIR is currently in revision. Its preprint is available in BioRxiv https://doi.org/10.1101/2020.08.07.240689, and the code is in GitHub https://github.com/camoragaq/BrumiR. This work was the core of the PhD of Carol Moraga Quinteros [37] funded by Conicyt and defended in September 2020. It is part of both a formal and an informal collaboration with 10.1.3.

8.4 Axis 3: (Co)Evolution

Cophylogeny

Phylogenetic tree reconciliation is the method of choice in analysing host-symbiont systems. Despite the many reconciliation tools that have been proposed in the literature, two main issues remain unresolved: (i) listing suboptimal solutions (*i.e.* whose score is "close" to the optimal ones) and (ii) listing only solutions that are biologically different "enough". The first issue arises because the optimal solutions are not always the ones biologically most significant; providing many suboptimal solutions as alternatives for the optimal ones is thus very useful. The second one is related to the difficulty to analyse a number of optimal solutions that is often exponential. We then proposed a method, that we called CAPYBARA for "equivalence ClAss enumeration of coPhylogenY event-BAsed ReconciliAtions" that addresses both of these problems in an efficient way. Furthermore, CAPYBARA includes a tool for visualising the solutions that significantly helps the user in the process of analysing the results. The source code, documentation, and binaries for all platforms are freely available at https://capybara-doc.readthedocs.io/. This work was published in *Bioinformatics* [27].

The problem of an efficient enumeration of equivalence classes or of one representative per class (without generating all the solutions), although identified as a need in many areas, has been addressed only for very few specific cases. In 2020, we started working on providing a general framework that solves this problem in polynomial delay in a wide variety of contexts, including optimisation ones that can be addressed by dynamic programming algorithms, and for certain types of equivalence relations between solutions. This theoretical work thus applies to a broad set of problems, among which phylogenetic tree reconciliation which initially motivated it. The work will be submitted in early 2021. It is already available in arXiv (https://arxiv.org/abs/2004.12143).

Theoretical aspects of cytoplasmic incompatibility

Another work, more purely theoretical but that was originally motivated by a question that is in some sense related to coevolution, concerned cytoplasmic incompatibility, CI for short. CI relates to the manipulation by the parasite *Wolbachia* of its host reproduction. Despite its widespread occurrence, the molecular basis of CI remains unclear and theoretical models have been proposed to understand the phenomenon. We considered in the work published in *Algorithms for Molecular Biology* [10] the quantitative Lock-Key model which currently represents a good hypothesis that is consistent with the data available. CI is in this case modelled as the problem of covering the edges of a bipartite graph with the minimum number of chain subgraphs. This problem was already known to be NP-hard, and we provide an exponential algorithm with a non-trivial complexity. It is frequent that depending on the dataset, there may be many optimal solutions which can be biologically quite different among them. To rely on a single optimal solution may therefore be problematic. To this purpose, we addressed the problem of enumerating (listing) all minimal chain subgraph covers of a bipartite graph and showed that it can be solved in quasi-polynomial time. Interestingly, in order to solve the above problems, we considered also the problem of enumerating all the maximal chain subgraphs of a bipartite graph and improved on the current results in the literature for the latter. Finally, to demonstrate the usefulness of our methods, we did show an application on a real dataset.

8.5 Axis 4: Health in general

Human

The essential part of the work on human health was focused on cancer and rare diseases. In the second case, such work concerns variants of three rare recessive developmental diseases, namely Taybi-Linder/MOPD1, Roifman and Lowry-Wood syndromes, and influenza A viruses. These works are presented in the context of (pan)genomics and transcriptomics in general, and more precisely of 8.2.

As concerns cancer, and more precisely breast cancer, two works were published in 2020.

The first one examined cancer cell plasticity and malignant progression, both of which remain poorly understood. In the paper published in *iScience* [25], we showed that the uncharacterised epigenetic factor chromodomain on Y-like 2 (CDYL2) is commonly over-expressed in breast cancer, and that high CDYL2 levels correlate with poor prognosis. Supporting a functional role for CDYL2 in malignancy, it positively regulated breast cancer cell migration, invasion, stem-like phenotypes, and epithelial-to-mesenchymal transition. CDYL2 regulation of these plasticity-associated processes depended on signalling via p65/NF-kB and STAT3. This, in turn, was downstream of CDYL2 regulation of MIR124 gene transcription. CDYL2 co-immunoprecipitated with G9a/EHMT2 and GLP/EHMT1 and regulated the chromatin enrichment of G9a and EZH2 at MIR124 genes. We then proposed that CDYL2 contributes to poor prognosis in breast cancer by recruiting G9a and EZH2 to epigenetically repress MIR124 genes, thereby promoting NF-kB and STAT3 signalling, as well as downstream cancer cell plasticity and malignant progression.

The second work on breast cancer concerned discovering disease signatures or subtypes through gene expression data analysis. Although this shows great promise, it is also prone to technical variation whose removal is essential to avoid spurious discoveries. Because this variation is not always known and can be confounded with biological signals, its removal is however a challenging task. In the paper published in *Communications Biology* [17], we provided a step-wise procedure and comprehensive analysis of the MINDACT microarray dataset. The MINDACT trial enrolled 6693 breast cancer patients and prospectively validated the gene expression signature MammaPrint for outcome prediction. The study also yielded a full-transcriptome microarray for each tumor. We showed for the first time in such a large dataset how technical variation can be removed while retaining expected biological signals.

Animal

Mycoplasma hyopneumoniae is the most costly pathogen for swine production. Although several studies have focused on the host-bacterium association, little is known about the changes in gene expression of swine cells upon infection. To improve our understanding of this interaction, we infected swine epithelial nptr cells with *M. hyopneumoniae* strain J to identify differentially expressed mRNAs and miRNAs. The levels of 1,268 genes and 170 miRNAs were significantly modified post-infection. Up-regulated mRNAs were enriched in genes related to redox homeostasis and antioxidant defense, known to be regulated by the transcription factor NRF2 in related species. Down-regulated mRNAs were enriched in genes associated with cytoskeleton and ciliary functions. Bioinformatic analyses suggested a correlation between changes in miRNA and mRNA levels, since we detected down-regulation of miRNAs predicted to target antioxidant genes and up-regulation of miRNAs targeting ciliary and cytoskeleton genes. Interestingly, most down-regulated miRNAs were detected in exosome-like vesicles suggesting that *M. hyopneumoniae* infection induced a modification of the composition of NPTr-released vesicles. Taken together, our data indicate that *M. hyopneumoniae* elicits an antioxidant response induced by NRF2 in infected cells. In addition, we propose that ciliostasis caused by this pathogen is partially explained by the down-regulation of ciliary genes. This work was published in *Scientific Reports* [22].

Others

Finally, we had in 2020 a work that is also related to health although in a less direct way. Indeed, in a paper published in *Journal of Operational Research* [26], we considered the problem of scheduling patients in allocated surgery blocks in a Master Surgical Schedule. We paid attention to both the available surgery blocks and the bed occupancy in the hospital wards. More specifically, large probabilities of overtime in each surgery block are undesirable and costly, while large fluctuations in the number of used beds requires extra buffer capacity and makes the staff planning more challenging. The stochastic nature of surgery durations and length of stay on a ward hinders the use of classical techniques. Transforming the stochastic problem into a deterministic problem does not result into practically feasible solutions. In this paper we developed a technique to solve the stochastic scheduling problem, whose primary objective it to minimise variation in the necessary bed capacity, while maximising the number of patients operated, and minimising the maximum waiting time, and guaranteeing a small probability of overtime in surgery blocks. The method starts with solving an Integer Linear Programming (ILP) formulation of the problem, and then simulation and local search techniques are applied to guarantee small probabilities of overtime and to improve upon the ILP solution. Numerical experiments applied to a Dutch hospital showed promising results.

9 Bilateral contracts and grants with industry

9.1 Bilateral grants with industry

Spock

- Title: characterization of hoSt-gut microbiota interactions and identification of key Players based on a unified reference for standardized quantitative metagenOmics and metaboliC analysis frameworK
- Industrial Partner: MaatPharma (Person responsible: Lilia Boucinha until 2019, Emmanuel Prestat from 2019).
- ERABLE participants: Marie-France Sagot (ERABLE coordinator and PhD main supervisor with Susana Vinga from IST, Lisbon, Portugal, as PhD co-supervisor), Marianne Borderes (beneficiary of the PhD scholarship in MaatPharma).
- Type: ANR Technology (2018-2021).
- Web page: `http://team.inria.fr/erable/en/projects/#anr-technology-spock`.

9.2 Informal Relations with Industry

Laurent Jacob works with Pendulum Therapeutics (previously Whole Biome) since 2019, with whom he signed an Non Disclosure Agreement and via whom he collaborates with Hector Roux de Bezieux, who is a PhD student in biostatistics at the University of California, Berkeley, USA, who is a computational biologist at the company.

10 Partnerships and cooperations

10.1 International initiatives

10.1.1 Inria Associate Team not involved in an Inria International Lab

Compasso

Title: COMmunity Perspective in the health sciences: Algorithms and Statistical approacheS for explOring it

Duration: 2018 - 2020

Coordinator: Marie-France Sagot

Partners:

- INESC-ID, Instituto Superior Técnico, Lisbonne (Portugal)

Inria contact: Marie-France Sagot

Summary: The final aim of this project is, through the development of mathematical models and algorithms, to start building links between species interactions and cancer/rare diseases, or more precisely, between infectious diseases and non infectious ones, whether they involve human or animals more in general. The main general questions that will be addressed are the following: (i) Are species interactions really a crucial factor on the development of at least some non infectious diseases as is suspected? (ii) If yes, could this disease be treated in a “non-aggressive" way by exploiting such species interactions? These are highly ambitious questions that will in the first three years be tackled through two angles. One concerns modelling and understanding the system biology of communities, and the second modelling and understanding the co-evolutionary aspects present in such communities. The first will in fact cover both synthetic communities and natural ones.

Capoeira

Title: Computational APproaches with the Objective to Explore intra and cross-species Interactions and their Role in All domains of life

Duration: 2020 - 2022

Coordinator: Marie-France Sagot

Partners:

- Instituto de Matemática e Estatistíca, Universidade de São Paulo (Brazil)

Inria contact: Marie-France Sagot

Summary: The CAPOEIRA project will cover theoretical computer science (essentially graph theory), mathematics (combinatorics, statistics, and probability), and the development of algorithms to address various biological questions, in particular, the intra and cross-species interactions, which have implications in all aspects of life sciences, including health, ecology, and environment. Two main general topics will be addressed, namely evolution/co-evolution, and biological network (graph/hypergraph) analysis and comparison. Both have already been explored by the partners. Some of the specific questions to be treated within each problem will thus represent a continuation of previous works. Each problem however also contains entirely new questions. Furthermore, the interaction with biologists within the project at both the modelling and validation steps is entirely new in the context of the past collaboration between the two partners. The first topic concerns better understanding and characterising the moment of speciation leading to new species on one hand, and on the other, how one set of species may influence the evolution of another. The second topic concerns metabolism on one hand, and (post-)transcriptional regulation on the other, with the post-transcriptional level involving also inference “from scratch” of the main actors, namely the non-coding RNAs and their targets, and the regulatory network they form. In the first two cases (of metabolism and transcriptional regulation), we will assume that the networks are already inferred albeit with possibly numerous missing and incorrect data. Finally, in the case of regulation, we will also consider the problem of inferring variants, notably related to alternative splicing, from a set of RNA-seq data using a de Bruijn graph approach. Overseeing these two main topics are the issues of knowledge representation and model revision that will also be addressed, which are crucial in life sciences, and notably in the context of post-transcriptional regulation by non-coding RNAs, for which the different actors, features, and overall mechanisms are constantly being questioned and revised.

10.1.2 Inria international partners

Informal international partners

Chile

Besides the collaboration with Elena Vidal from the Universidad Mayor, Santiago, mentioned below, we have informal collaborations with Rodrigo Gutiérrez from the Universidad Catolica of Santiago who was co-supervisor of Carol Moraga Quinteros with M.-F. Sagot, as well as with Vicente Acuña, who is Scientist at the Centro de Modelamiento Matemático (CMM), at Santiago.

10.1.3 Participation in other international programs

Brazil

Ahimsa

- Title: Alternative approacH to Investigating and Modelling Sickness and heAlth
- Coordinators: M.-F. Sagot (ERABLE), A. Ávila (Instituto de Biologia Molecular do Paraná – Fiocruz-PR, Curitiba, Paraná, Brazil)
- ERABLE participant(s): M. Ferrarini, A. Mary, S. Mucha, M.-F. Sagot, B. Sinaimeri
- Type: Capes-Cofecub (2020-2022)
- Web page: https://team.inria.fr/erable/en/projects/capes-cofecub-project-ahimsa/

Fapesp-UdL

- Title: Graph/Hypergraph (spectral) analysis to compare metabolic networks of pathogenic *Trypanosoma* sp.
- Coordinators: M.-F. Sagot (ERABLE), A. Fujita (University of São Paulo (USP), São Paulo, Brazil)
- ERABLE participant(s): M. Ferrarini, V. Lacroix, A. Mary, M.-F. Sagot, B. Sinaimeri
- Type: Fapesp-UcL (2020-2021)
- Web page: Not available

Chile

ERABLE participates in the project Network for Organismal Interactions Research (NOIR) funded by Conicyt in Chile within the call International Networking between Research Centers. The project started in 2019 and will last until the end of 2020. The coordinator on the Chilean side is Elena Vidal from the Universidad Mayor, Santiago, Chile, and the Erable participants are Carol Moraga Quinteros, Mariana Ferrarini and Marie-France Sagot.

10.2 International research visitors

10.2.1 Visits of international scientists

Due to the Covid-19, a number of planned visits, notably from Brazil, Chile, Portugal, etc, had to be cancelled. The only one that took place in the first two months of 2020 concerned Nuno Mira, our collaborator from the Instituto Superior Técnico (IST), Lisbon, who was cominh for a Sabbatical until the middle of 2020. Nuno managed to get back to his family in Lisbon just as the first confinement was going to start in mid-March. He has been negociating with IST to postpone his Sabbatical, and thus his long visit to us to 2021, or maybe even later depending on how the situation evolves.

10.2.2 Visits to international teams

Research stays abroad Blerina Sinaimeri did a 12-month Sabbatical at Luiss University in Rome, Italy, where a member of ERABLE, Giuseppe Italiano, is Full Professor, thus interacting with him and also with another ERABLE team member, Alberto Marchetti-Spaccamela who is Full Professor at Sapienza University in Rome. Blerina's stay actually started in July 2019, and will be extended until the end of January 2021. Starting from February 2021, Blerina will continue at Luiss University as an Associate Professor, with a 3-years Detachment from Inria. Blerina will continue to be member of ERABLE as is Giuseppe Italiano, Alberto Marchetti-Spaccamela and other researchers at the University of Pisa in Italy.

10.3 European initiatives

10.3.1 FP7 & H2020 Projects

Two ERABLE members in the Netherlands, Solon Pissis and Leen Stougie, participate in an H2020 MSCA-RISE project (2020-2022) called Pangaia (Pan-genome Graph Algorithms and Data Integration) coordinated by Paola Bonizzoni, University of Milan, Italy.

Furthermore, in 2020, an H2020 Twinning project was accepted that is coordinated by INESC-ID, Instituto Superior Técnico, Lisbon, Portugal and of which ERABLE is a partner (the coordinator in France is Marie-France Sagot). The project focuses on Computational Biology, a strongly interdisciplinary area that combines Computer Science, Algorithms, Mathematics, Probability and Statistics, Machine Learning, Molecular Biology and Medicine. The project consortium is composed of INESC-ID (Coordinator), the National Institute for Research in Digital Science and Technology (Inria) through the Erable team, the Swiss Federal Institute of Technology (ETH Zürich) in Switzerland and the European Molecular Biology Laboratory (EMBL) in Germany. The main goal of the project is to intensify, increase and consolidate the research in Computational Biology carried out at INESC-ID in partnership with the European partner institutions.

Due to the Covid-19, the start of this project was delayed until January 1st, 2021. It will last until the end of 2023, unless it is extended due to the fact that some of the planned initiatives for the first year may not be realisable, once again because of the Covid-19.

In the same way, another H2020 project, in this case an ITN with acronym Alpaca that involves members of ERABLE has been accepted in 2020 but will start only in 2021. Two members of ERABLE will host a PhD student in their institutions, namely Solon Pissis at CWI and Nadia Pisanti at the University of Pisa. Other members of ERABLE will be involved in Alpaca.

10.3.2 Collaborations with major European organizations

By itself, ERABLE is built from what initially were collaborations with some major European Organisations (CWI, Sapienza University of Rome, Universities of Florence and Pisa, Free University of Amsterdam) and then became a European Inria Team.

10.4 National initiatives

10.4.1 ANR

Aster

- Title: Algorithms and Software for Third gEneration Rna sequencing
- Coordinator: Hélène Touzet, University of Lille and CNRS.
- ERABLE participants: Vincent Lacroix (ERABLE coordinator), Audric Cologne, Eric Cumunel, Alex di Genova, Leandro I. S. de Lima, Arnaud Mary, Marie-France Sagot, Camille Sessegolo, Blerina Sinaimeri.
- Type: ANR (2016-2020).
- Web page: `http://bioinfo.cristal.univ-lille.fr/aster/`.

Fast-Big

- Title: Efficient Statistical Testing for high-dimensional Models: application to Brain Imaging and Genomics
- Coordinator: B. Thirion
- ERABLE participant(s): L. Jacob, A. Villié
- Type: ANR (2018-2022)
- Web page: https://anr.fr/Project-ANR-17-CE23-0011

GrR

- Title: Graph Reconfiguration
- Coordinator: N. Bousquet
- ERABLE participant(s): A. Mary
- Type: ANR JCJC (2019-2021)
- Web page: Not available

Green

- Title: Deciphering host immune gene regulation and function to target symbiosis disturbance and endosymbiont control in insect pests
- Coordinator: A. Heddi
- ERABLE participant(s): M.-F. Sagot, C. Vieira
- Type: ANR (2018-2021)
- Web page: https://www.insa-lyon.fr/fr/green

Hmicmac

- Title: Host-microbiota co-adaptations: mechanisms and consequences
- Coordinator: F. Vavre
- ERABLE participant(s): F. Vavre
- Type: ANR PRC (2017-2020)
- Web page: Not available

Resist

- Title: Rapid Evolution of Symbiotic Interactions in response to STress: processes and mechanisms
- Coordinator: N. Kremer
- ERABLE participant(s): F. Vavre
- Type: ANR JCJC (2017-2020)
- Web page: Not available

Swing

- Title: Worldwide invasion of the Spotted WING Drosophila: Genetics, plasticity and evolutionary potential
- Coordinator: P. Gibert
- ERABLE participant(s): C. Vieira
- Type: ANR PCR (2016-2020)
- Web page: Not available

U4atac-brain

- Title: Rôle de l'épissage mineur dans le développement cérébral
- Coordinator: Patrick Edery, Centre de Recherche en Neurosciences de Lyon.
- ERABLE participants: Vincent Lacroix (ERABLE coordinator), Audric Cologne.
- Type: ANR (2018-2021)
- Web page: Not available

10.4.2 Idex

Micro-be-have

- Title: Microbial Impact on insect behaviour: from niche and partner selection to the development of new control methods for pests and disease vectors
- Coordinator: F. Vavre
- ERABLE participant(s): F. Vavre
- Type: AO Scientific Breakthrough (2018-2021)
- Web page: Not available

10.4.3 Others

Notice that were included here national projects of our members from Italy and the Netherlands when these have no other partners than researchers from the same country.

AHeAD

- Title: efficient Algorithms for HArnessing networked Data
- Coordinator: G. Italiano
- ERABLE participant(s): R. Grossi, G. Italiano
- Type: MUIR PRIN, Italian Ministry of Education, University and Research (2019-2022)
- Web page: https://sites.google.com/view/aheadproject

Networks

- Title: Networks
- Coordinator: Michel Mandjes, University of Amsterdam
- ERABLE participant(s): S. Pissis, L. Stougie
- Type: NWO Gravity Program (2014-2024)
- Web page: https://www.thenetworkcenter.nl/

Optimal

- Title: Optimization for and with Machine Learning
- Coordinator: Dick den Hertog
- ERABLE participant(s): L. Stougie
- Type: NWO ENW-Groot Program
- Web page: Not available

11 Dissemination

11.1 Promoting scientific activities

11.1.1 Scientific events: organisation

General chair, scientific chair

- Giuseppe Italiano is member of the Steering Committee of the *Workshop on Algorithm Engineering and Experimentation (ALENEX)*, of the International Colloquium on Automata, Languages and Programming (ICALP), and of the Workshop/Symposium on Experimental Algorithms (SEA).
- Alberto Marchetti-Spaccamela is a member of the Steering committee of *Workshop on Graph Theoretic Concepts in Computer Science (WG)*), and of *Workshop on Algorithmic Approaches for Transportation Modeling, Optimization, and Systems (ATMOS)*.
- Arnaud Mary is member of the Steering Committee of *Workshop on Enumeration Problems and Applications (WEPA)*.
- Marie-France Sagot is member of the Steering Committee of *European Conference on Computational Biology (ECCB)*, *International Symposium on Bioinformatics Research and Applications (ISBRA)*, and *Workshop on Enumeration Problems and Applications (WEPA)*.

Member of the organizing committees

- Roberto Grossi was chair of the Organizing Committee of *ALGO*.
- Laurent Jacob was a member of the Organizing Committee of the Colloquium "Prospective en sciences des données, intelligence artificielle et biologie" organised by the CNRS INSB.
- Nadia Pisanti was a member of the Organizing Committee of *ALGO*.
- Marie-France Sagot is co-chair of the joint conferences *ISMB/ECCB* that will take place in Lyon in July 2021 and whose organisation started more than one year before.

11.1.2 Scientific events: selection

Chair of conference program committees

- Nadia Pisanti was a chair of the Program Committee of *WABI*.

Member of the conference program committees

- Roberto Grossi was a member of the Program Committee of *CPM*, *iABC*, *SWAT*, and *WWW*.
- Giuseppe Italiano was a member of the Program Committee of *APF*, *ATMOS*, and *CIAC*.
- Laurent Jacob was a member of the Program Committee of *JOBIM*.
- Arnaud Mary was a member of the Program Committee of *WEPA*.
- Nadia Pisanti was a member of the Program Committee of *BIOINFORMATICS*, *iABC*, *ICTCS*, *ISBRA*, *RDAAPS*, *SOFSEM*, *SEA*, and *SPIRE*.
- Solon Pissis was a member of the Program Committee of *WABI*.
- Marie-France Sagot was a member of the Program Committee of *LAGOS*, *RECOMB*, *RecombCG*, *WABI*, and *WEPA*.

Reviewer Members of ERABLE have reviewed papers for a number of workshops and conferences including: *CIAC*, *CPM*, *ISBRA*, *ISMB*, *LAGOS*, *MLCB*, *NeurIPS*, *RECOMB*, *WEPA*, *WABI*.

11.1.3 Journal

Member of the editorial boards

- Roberto Grossi is member of the Editorial Board of *Theory of Computing Systems (TOCS)* and pf *RAIRO – Theoretical Informatics and Applications*.
- Giuseppe Italiano is member of the Editorial Board of *Algorithmica* and *Theoretical Computer Science*.
- Vincent Lacroix is recommender for *Peer Community in Genomics*, see `https://genomics.peercommunityin.org/`.
- Alberto Marchetti-Spaccamela is member of the Editorial Board of *Theoretical Computer Science*.
- Nadia Pisanti is since 2017 member of Editorial Board of *Network Modeling Analysis in Health Informatics and Bioinformatics*.
- Marie-France Sagot is member of the Editorial Board of *BMC Bioinformatics*, *Algorithms for Molecular Biology*, and *Lecture Notes in BioInformatics*.
- Leen Stougie is member of the Editorial Board of *AIMS Journal of Industrial and Management Optimization*.
- Cristina Vieira is Executive Editor of *Gene*, and since 2014 member of the Editorial Board of *Mobile DNA*.

Reviewer - reviewing activities Members of ERABLE have reviewed papers for a number of journals including: *Theoretical Computer Science*, *Algorithmica*, *Algorithms for Molecular Biology*, *Bioinformatics*, *BMC Bioinformatics*, *Genome Biology*, *Genome Research*, *IEEE/ACM Transactions in Computational Biology and Bioinformatics (TCBB)*, *Machine Learning*, *Molecular Biology and Evolution*, *Nucleic Acid Research*.

11.1.4 Invited talks

- Laurent Jacob: Keynote talk on "Learning with pangenomes" at Statistical Methods for Post Genomic Data, Jan 2020; Keynote talk on "Learning from sequences with convolutional kernel networks" at Mini-symposium Deep Learning of Jobim 2020, Jun 2020; Invited talk on "Statistical pangenomics" at Seminar of the Biostatistics and Epidemiology Service at Institut Gustave Roussy, Jun 2020;Invited talk on “Learning with pangenomes”, Seminar of the Inrae-MIAT, Toulouse, Sep 2020.

11.1.5 Scientific expertise

Hubert Charles was until this year director of the Biosciences Department of the Insa-Lyon and co-director of studies of the “Bioinformatique et Modélisation (BIM)” track.

Roberto Grossi is member of the International Olympiad in Informatics.

Giuseppe Italiano is Vice-President of the European Association for Theoretical Computer Science (EATCS) and Director of the Master of Science in Data Science and Management, LUISS University, Rome, besides having a number of other responsabilities at LUISS. He is also member of the Advisory Board of MADALGO - Center for MAssive Data ALGOrithmics, Aarhus, Denmark.

Laurent Jacob is alternate member of the “Conseil National des Universités” (CNU) 26 (“Applied Mathematics and Applications of Mathematics”). He is in the "Conseil Pédagogique et d’Orientation" and participated to the recruitment committee for an Associate Professor at Polytech Grenoble.

Nadia Pisanti is since November 1st 2017 member of the Board of the PhD School in Data Science (University of Pisa jointly with Scuola Normale Superiore Pisa, Scuola S. Anna Pisa, IMT Lucca).

Marie-France Sagot is member of the Advisory Board of CWI, Amsterdam, the Netherlands. In 2020, she was part of the ERC Consolidator Panel for LS2 and a member of the Review Committee for the Human Frontier Science Program.

Blerina Sinaimeri was member of the Inria National Commission for the recruitment of junior researchers, CRCN and ISFP, in 2020.

Leen Stougie is since April 2017 Leader of the Life Science Group at CWI. He is member of the General Board of the Dutch Network on the Mathematics of Operations Research (Landelijk Netwerk Mathematische Besliskunde (LNMB)), member of the Management Team of the Gravity project Networks, and member of the Gijs de Leve Award 2021 committee.

Alain Viari is member of a number of scientific advisory boards (IRT (Institut de Recherche Technologique) BioAster; Centre Léon Bérard). He also coordinates together with J.-F. Deleuze (CNRGH-Evry) the Research & Development part (CRefIX) of the “Plan France Médecine Génomique 2025”.

Fabrice Vavre is President of the Section 29 of the CoNRS8.

Cristina Vieira is member of the “Conseil National des Universités” (CNU) 67 (“Biologie des Populations et Écologie”), and since 2017 member of the “Conseil de la Faculté des Sciences et Technologies (FST)” of the University Lyon 1.

11.2 Teaching - Supervision - Juries

11.2.1 Teaching

France The members of ERABLE teach both at the Department of Biology of the University of Lyon (in particular within the BISM (BioInformatics, Statistics and Modelling) specialty, and at the department of Bioinformatics of the Insa (National Institute of Applied Sciences). Cristina Vieira is responsible for the Master Biodiversity, Ecology and Evolution (https://www.bee-lyon-univ.fr/). She teaches genetics 192 hours per year at the University and at the ENS-Lyon. Hubert Charles was until this year responsible for the Master of Modelling and Bioinformatics (BIM) at the Insa of Lyon (http://biosciences.insa-lyon.fr/). He teaches 192 hours per year in statistics and biology. Vincent Lacroix is responsible for the M1 master in bioinformatics (https://www.bioinfo-lyon.fr/) and of the following courses (L3: Advanced Bioinformatics, M1: Methods for Data Analysis in Genomics, M1: Methods for Data Analysis in Transcriptomics, M1: Bioinformatics Project, M2: Ethics). He taught 192 hours in 2020. Arnaud Mary is responsible for three courses of the Bioinformatics Curriculum at the University (L2: Introduction to Bioinformatics and Biostatistics, M1: Object Oriented Programming, M2: new course on Advanced

Algorithms for Bioinformatics) and one at Insa (Discrete Mathematics). He taught 198 hours in 2020. Blerina Sinaimeri Blerina Sinaimeri taught 12H hours in 2020 on graph algorithms for the M1 students of the Master in Bioinformatics. Fabrice Vavre taught 20h at the Master level.

The ERABLE team regularly welcomes M1 and M2 interns from the bioinformatics Master.

All French members of the ERABLE team are affiliated to the doctoral school E2M2 (Ecology-Evolution-Microbiology-Modelling, http://e2m2.universite-lyon.fr/).

Italy & The Netherlands Italian researchers teach between 90 and 140 hours per year, at both the undergraduate and at the Master levels. The teaching involves pure computer science courses (such as Programming foundations, Programming in C or in Java, Computing Models, Distributed Algorithms) and computational biology (such as Algorithms for Bioinformatics).

Dutch researchers teach between 60 and 100 hours per year, again at the undergraduate and Master levels, in applied mathematics (*e.g.* Operational Research, Advanced Linear Programming), machine learning (Deep Learning) and computational biology (*e.g.* Biological Network Analysis, Algorithms for Genomics).

11.2.2 Supervision

The following PhDs were defended in ERABLE in 2020:

- Dexiong Chen, Université Grenoble Alpes, Rome (co-supervisors: Julien Mairal, Laurent JAcob), Dec 2020.
- Aikaterini Karanasiou, University Tor Vergata, Rome (co-supervisor: Giuseppe Italiano), Apr 2020.
- Shima Moghtasedi, University of Pisa (supervisor: Roberto Grossi), Apr 2020.
- Carol Moraga Quinteros, University of Lyon 1 (funded by Conicyt Chile, co-supervisors: Rodrigo Gutierrez – Catholic University of Chile, Marie-France Sagot), Sep 2020.
- Irene Ziska, University Lyon 1 (funded by Inria Cordi-S, co-supervisors: Susana Vinga – Instituto Superior Técnico at Lisbon; Marie-France Sagot), Nov 2020.

The following are the PhDs in progress:

- Giulia Bernardini, University Milan-Bicocca (co-supervisor: Nadia Pisanti).
- Marianne Borderes, University Lyon 1 (funded by ANR Technology Spock, co-supervisors: Susana Vinga – Instituto Superior Técnico at Lisbon; Marie-France Sagot).
- Nicolas Homberg, Inra, Inria & University of Lyon 1 (funded by Inra & Inria, co-supervisors: Christine Gaspin at Inra; Marie-France Sagot).
- Francesca Lizzi, Scuola Normale Superiore, Pisa (co-supervisor: Nadia Pisanti)
- Giulia Punzi, University of Pisa (supervisor: Roberto Grossi)
- Camille Sessegolo, University of Lyon 1 (funded by ANR Aster; co-supervisors: Vincent Lacroix, Arnaud Mary).
- Michelle Sweering, CWI (co-supervisors: Solon Pissis and Leen Stougie).
- Luca Versari, University of Pisa (supervisor: Roberto Grossi; Software Engineer at Google Research, Zürich, since 2019)
- Antoine Villié, Université Lyon 1(funded by ANR Fast-Big, supervisor: Laurent Jacob).
- Yishu Wang, University Lyon 1 (funded by Ministère de l'Enseignement supérieur, de la Recherche et de l'Innovation, co-supervisors: Mário Figueiredo – Instituto Superior Técnico at Lisbon; Marie-France Sagot; Blerina Sinaimeri).

11.2.3 Juries

The following are the PhD or HDR juries to which members of ERABLE participated in 2020.

- Laurent Jacob: External reviewer of the PhD of Guillaume Gautreau, supervised by Claudine Médigue and David Vallenet, Université d'Évry Val d'Essone, Mar 2020; External reviewer of the PhD of Olga Permiakova, supervised by Thomas Burger, Université Grenobles Alpes which should have taken place in Dec 2020 but was postponed to May 2021 due to the Covid-19.

- Blerina Sinaimeri: Participant in the PhD committee of Aikaterini Karanasiou, supervised by Giuseppe Italiano, University Tor Vergata, Rome, Apr 2020.

- Leen Stougie: Chair of the PhD committee of Amir Shabani, supervised by Wout Dullaert, and Gabor Maroti, Free University of Amsterdam, Jul 2020; Principal Opponent of the PhD of Alexander Birx, supervised by Yann Disser, Technical University of Darmstadt, Oct 2020; External reviewer of the PhD of Madelon de Kemp, supervised by Michel Mandjes and Neil Olver, University of Amsterdam, Dec 2020.

12 Scientific production

12.1 Publications of the year

International journals

[1] M. Alzamel, L. A. K. Ayad, G. Bernardini, R. Grossi, C. S. Iliopoulos, N. Pisanti, S. P. Pissis and G. Rosone. 'Comparing Degenerate Strings'. In: *Fundamenta Informaticae* 175 (28th Sept. 2020), pp. 41–58. DOI: 10.3233/fi-2020-1947. URL: https://hal.inria.fr/hal-03085839.

[2] R. Andrade, M. Doostmohammadi, J. Santos, M.-F. Sagot, N. P. Mira and S. Vinga. 'MOMO - multi-objective metabolic mixed integer optimization: application to yeast strain engineering'. In: *BMC Bioinformatics* 21.1 (Dec. 2020), pp. 1–13. DOI: 10.1186/s12859-020-3377-1. URL: https://hal.inria.fr/hal-02490353.

[3] U. Ashraf, C. Benoit-Pilven, V. Navratil, C. Ligneau, G. Fournier, S. Munier, O. Sismeiro, J.-Y. Coppée, V. Lacroix and N. Naffakh. 'Influenza virus infection induces widespread alterations of host cell splicing'. In: *NAR Genomics and Bioinformatics* 2.4 (21st Nov. 2020). DOI: 10.1093/nargab/lqaa095. URL: https://hal.archives-ouvertes.fr/hal-03021806.

[4] L. A. Ayad, G. Bernardini, R. Grossi, C. S. Iliopoulos, N. Pisanti, S. P. Pissis and G. Rosone. 'Longest property-preserved common factor: A new string-processing framework'. In: *Theoretical Computer Science* 812 (Apr. 2020), pp. 244–251. DOI: 10.1016/j.tcs.2020.02.012. URL: https://hal.inria.fr/hal-02956135.

[5] A. C. Baller, M. Van Ee, M. Hoogeboom and L. Stougie. 'Complexity of inventory routing problems when routing is easy'. In: *Networks* 75.2 (2020), pp. 113–123. DOI: 10.1002/net.21908. URL: https://hal.inria.fr/hal-02422721.

[6] A. Bénard, F. Vavre and N. Kremer. 'Stress & Symbiosis: Heads or Tails?' In: *Frontiers in Ecology and Evolution* (9th June 2020), pp. 1–9. DOI: 10.3389/fevo.2020.00167. URL: https://hal-cnrs.archives-ouvertes.fr/hal-02908996.

[7] C. Benoit-Pilven, A. Besson, A. Putoux, C. Benetollo, C. Saccaro, J. Guguin, G. Sala, A. Cologne, M. Delous, G. Lesca, R. A. Padgett, A.-L. Leutenegger, V. Lacroix, P. Edery and S. Mazoyer. 'Clinical interpretation of variants identified in RNU4ATAC, a non-coding spliceosomal gene'. In: *PLoS ONE* 15.7 (6th July 2020), e0235655. DOI: 10.1371/journal.pone.0235655. URL: https://www.hal.inserm.fr/inserm-02915106.

[8] G. Bernardini, H. Chen, A. Conte, R. Grossi, G. Loukides, N. Pisanti, S. P. Pissis, G. Rosone and M. Sweering. 'Combinatorial Algorithms for String Sanitization'. In: *ACM Transactions on Knowledge Discovery from Data (TKDD)* 15.1 (7th Dec. 2020), pp. 1–34. DOI: 10.1145/3418683. URL: https://hal.inria.fr/hal-03085838.

[9] M. Borderes, C. Gasc, E. Prestat, M. G. Ferrarini, S. Vinga, L. Boucinha and M.-F. Sagot. 'A comprehensive evaluation of binning methods to recover human gut microbial species from a non-redundant reference gene catalog'. In: *NAR Genomics and Bioinformatics* 3.1 (2021). DOI: 10.1093/nargab/lqab009. URL: https://hal.inria.fr/hal-03157241.

[10] T. Calamoneri, M. Gastaldello, A. Mary, M.-F. Sagot and B. Sinaimeri. 'Algorithms for the quantitative Lock/Key model of cytoplasmic incompatibility'. In: *Algorithms for Molecular Biology* (2020), pp. 1–16. DOI: 10.1186/s13015-020-00174-1. URL: https://hal.inria.fr/hal-02917569.

[11] M. Castro, C. Goubert, F. Monteiro, C. Vieira and C. M. A. Carareto. 'Homology-Free Detection of Transposable Elements Unveils Their Dynamics in Three Ecologically Distinct Rhodnius Species'. In: *Genes* 11.2 (Feb. 2020), p. 170. DOI: 10.3390/genes11020170. URL: https://hal.archives-ouvertes.fr/hal-02487180.

[12] F. Catez, N. Dalla-Venezia, J.-J. Diaz, B. Dubois, A. Ferrari, B. Guyot, V. Marcel, M. Ouzounova, R. M. Pommier, A. Viari and P. Mehlen. 'Actualités en recherche en oncologie : l'essentiel du 4e Symposium International 2019 du Centre de Recherche en Cancérologie de Lyon'. In: *Bulletin du Cancer* 107.1 (Jan. 2020), pp. 136–140. DOI: 10.1016/j.bulcan.2019.12.002. URL: https://hal.archives-ouvertes.fr/hal-03004269.

[13] J. Consuegra, T. Grenier, P. Baa-Puyoulet, I. Rahioui, H. Akherraz, H. Gervais, N. Parisot, P. Da Silva, H. Charles, F. Calevro and F. Leulier. 'Drosophila-associated bacteria differentially shape the nutritional requirements of their host during juvenile growth'. In: *PLoS Biology* 18.3 (20th Mar. 2020), e3000681. DOI: 10.1371/journal.pbio.3000681. URL: https://hal.archives-ouvertes.fr/hal-02919023.

[14] A. Conte, D. De Sensi, R. Grossi, A. Marino and L. Versari. 'Truly Scalable K-Truss and Max-Truss Algorithms for Community Detection in Graphs'. In: *IEEE Access* 8 (2020), pp. 139096–139109. DOI: 10.1109/ACCESS.2020.3011667. URL: https://hal.inria.fr/hal-02956066.

[15] M. G. Ferrarini, L. M. Nisimura, R. M. B. M. Girard, M. B. Alencar, M. S. I. Fragoso, C. A. Araújo-Silva, A. D. A. Veiga, A. P. R. Abud, S. C. Nardelli, R. C. Vommaro, A. M. Silber, M. France-Sagot and A. R. Ávila. 'Dichloroacetate and Pyruvate Metabolism: Pyruvate Dehydrogenase Kinases as Targets Worth Investigating for Effective Therapy of Toxoplasmosis'. In: *MSphere* 6.1 (24th Feb. 2021). DOI: 10.1128/mSphere.01002-20. URL: https://hal.inria.fr/hal-03104886.

[16] A. D. Genova, E. Buena-Atienza, S. Ossowski and M.-F. Sagot. 'Efficient hybrid de novo assembly of human genomes with WENGAN'. In: *Nature Biotechnology* (14th Dec. 2020). DOI: 10.1038/s41587-020-00747-w. URL: https://hal.inria.fr/hal-03065904.

[17] L. Jacob, A. Witteveen, I. Beumer, L. Delahaye, D. Wehkamp, J. Van Den Akker, M. Snel, B. Chan, A. Floore, N. Bakx, G. Brink, C. Poncet, J. Bogaerts, M. Delorenzi, M. Piccart, E. Rutgers, F. Cardoso, T. Speed, L. Van 't Veer and A. Glas. 'Controlling technical variation amongst 6693 patient microarrays of the randomized MINDACT trial'. In: *Communications Biology* 3 (27th July 2020). DOI: 10.1038/s42003-020-1111-1. URL: https://hal-cnrs.archives-ouvertes.fr/hal-02990043.

[18] M. Kapun, M. Barrón, F. Staubach, D. Obbard, R. A. W. Wiberg, J. Vieira, C. Goubert, O. Rota-Stabelli, M. Kankare, M. Bogaerts-márquez, A. A. Haudry, L. Waidele, I. Kozeretska, E. Pasyukova, V. Loeschcke, M. Pascual, S. Serga, C. Montchamp-Moreau, J. Abbott, P. Gibert, D. Porcelli, N. Posnien, A. Sánchez-Gracia, S. Grath, É. Sucena, A. Bergland, M. P. G. Guerreiro, B. S. Onder, E. Argyridou, L. Guio, M. F. Schou, B. Deplancke, C. Vieira, M. Ritchie, B. Zwaan, E. Tauber, D. Orengo, E. Puerma, M. Aguadé, P. Schmidt, J. Parsch, A. Betancourt, T. Flatt and J. González. 'Genomic Analysis of European Drosophila melanogaster Populations Reveals Longitudinal Structure, Continent-Wide Selection, and Previously Unknown DNA Viruses'. In: *Molecular Biology and Evolution* 37.9 (1st Sept. 2020), pp. 2661–2678. DOI: 10.1093/molbev/msaa120. URL: https://hal.archives-ouvertes.fr/hal-02957899.

[19] V. Mérel, M. Boulesteix, M. Fablet and C. Vieira. 'Transposable elements in Drosophila'. In: *Mobile DNA* 11.1 (Dec. 2020). DOI: 10.1186/s13100-020-00213-z. URL: https://hal.archives-ouvertes.fr/hal-02957906.

[20] A. Monti and B. Sinaimeri. 'String factorisations with maximum or minimum dimension'. In: *Theoretical Computer Science* 842 (Nov. 2020), pp. 65–73. DOI: 10.1016/j.tcs.2020.07.029. URL: https://hal.inria.fr/hal-02955643.

[21] M. Mourdas, N. T.-M. Dang, Y. Ogyama, N. Burlet, B. Mugat, M. Boulesteix, V. Mérel, P. Veber, J. Salces-Ortiz, D. Severac, A. Pélisson, C. Vieira, F. Sabot, M. Fablet and S. Chambeyron. 'A transposon story : from TE content to TE dynamic invasion of Drosophila genomes using the single-molecule sequencing technology from Oxford Nanopore'. In: *Cells* (2020). DOI: 10.3390/cells9081776. URL: https://hal.archives-ouvertes.fr/hal-02905128.

[22] S. G. Mucha, M. G. Ferrarini, C. Moraga, A. D. Genova, L. Guyon, F. Tardy, S. Rome, M.-F. Sagot and A. Zaha. 'Mycoplasma hyopneumoniae J elicits an antioxidant response and decreases the expression of ciliary genes in infected swine epithelial cells'. In: *Scientific Reports* 10.13707 (2020). DOI: 10.1038/s41598-020-70040-y. URL: https://hal.inria.fr/hal-02916844.

[23] N. Prezza, N. Pisanti, M. Sciortino and G. Rosone. 'Variable-order reference-free variant discovery with the Burrows-Wheeler Transform'. In: *BMC Bioinformatics* 21.S8 (Sept. 2020). DOI: 10.1186/s12859-020-03586-3. URL: https://hal.inria.fr/hal-02957620.

[24] T. Pusa, M. G. Ferrarini, R. Andrade, A. Mary, A. Marchetti-Spaccamela, L. Stougie and M.-F. Sagot. 'MOOMIN – Mathematical explOration of 'Omics data on a MetabolIc Network'. In: *Bioinformatics* 36.2 (2020), pp. 514–523. DOI: 10.1093/bioinformatics/btz584. URL: https://hal.inria.fr/hal-02284835.

[25] M. Siouda, A. D. Dujardin, L. Barbollat-Boutrand, M. Mendoza-Parra, B. Gibert, M. Ouzounova, J. Bouaoud, L. Tonon, M. Robert, J.-P. Foy, V. Lavergne, S. Manie, A. Viari, A. Puisieux, G. Ichim, H. Gronemeyer, P. Saintigny and P. Mulligan. 'CDYL2 Epigenetically Regulates MIR124 to Control NF-kB/STAT3-Dependent Breast Cancer Cell Plasticity'. In: *iScience* 23.6 (June 2020), p. 101141. DOI: 10.1016/j.isci. URL: https://hal.inria.fr/hal-02956108.

[26] A. Van Den Broek D'Obrenan, A. Ridder, D. Roubos and L. Stougie. 'Minimizing bed occupancy variance by scheduling patients under uncertainty'. In: *European Journal of Operational Research* 286.1 (Oct. 2020), pp. 336–349. DOI: 10.1016/j.ejor.2020.03.026. URL: https://hal.inria.fr/hal-02971122.

[27] Y. Wang, A. Mary, M.-F. Sagot and B. Sinaimeri. 'Capybara: equivalence ClAss enumeration of coPhylogenY event-BAsed ReconciliAtions'. In: *Bioinformatics* 36.14 (15th Aug. 2020), pp. 4197–4199. DOI: 10.1093/bioinformatics/btaa498. URL: https://hal.inria.fr/hal-02917341.

International peer-reviewed conferences

[28] V. Acuña, L. I. Soares De Lima, G. F. Italiano, L. P. Sciarria, M.-F. Sagot and B. Sinaimeri. 'A Family of Tree-Based Generators for Bubbles in Directed Graphs'. In: IWOCA 2020 - 31st International Workshop on Combinatorial Algorithms. Vol. 12126. Lecture Notes in Computer Science. Bordeaux, France, 29th May 2020, pp. 17–29. DOI: 10.1007/978-3-030-48966-3_2. URL: https://hal.inria.fr/hal-02971154.

[29] M. Alzamel, A. Conte, S. Denzumi, R. Grossi, C. S. Iliopoulos, K. Kurita and K. Wasa. 'Finding the Anticover of a String'. In: CPM 2020 - 31st Annual Symposium on Combinatorial Pattern Matching. Vol. 161. Leibniz International Proceedings in Informatics, LIPIcs. Copenhagen, Denmark, 17th June 2020, pp. 1–11. DOI: 10.4230/LIPIcs.CPM.2020.2. URL: https://hal.inria.fr/hal-02957658.

[30] J. A. Baaijens, L. Stougie and A. Schönhuth. 'Strain-Aware Assembly of Genomes from Mixed Samples Using Flow Variation Graphs'. In: RECOMB 2020 - 24th International Conference on Research in Computational Molecular Biology. Vol. 12074. Lecture Notes in Computer Science. Padova (Virtual), Italy, 21st Apr. 2020, pp. 221–222. DOI: 10.1007/978-3-030-45257-5_14. URL: https://hal.inria.fr/hal-02955692.

[31] G. Bernardini, H. Chen, A. Conte, R. Grossi, G. Loukides, N. Pisanti, S. Pissis and G. Rosone. 'String Sanitization: A Combinatorial Approach'. In: Joint European Conference on Machine Learning and Knowledge Discovery in Databases. Vol. 11906. Lecture Notes in Computer Science. Würzburg, Germany, 30th Apr. 2020, pp. 627–644. DOI: 10.1007/978-3-030-46150-8_37. URL: https://hal.inria.fr/hal-03085832.

[32] G. Bernardini, H. Chen, G. Loukides, N. Pisanti, S. P. Pissis, L. Stougie and M. Sweering. 'String Sanitization Under Edit Distance'. In: CPM 2020 - 31st Annual Symposium on Combinatorial Pattern Matching. Vol. 161. Leibniz International Proceedings in Informatics (LIPIcs). Copenhagen, Denmark, 17th June 2020, pp. 1–14. DOI: 10.4230/LIPIcs.CPM.2020.7. URL: https://hal.inria.fr/hal-02957647.

[33] P. Charalampopoulos, S. P. Pissis, J. Radoszewski, T. Waleń and W. Zuba. 'Unary Words Have the Smallest Levenshtein k-Neighbourhoods'. In: 31st Annual Symposium on Combinatorial Pattern Matching (CPM 2020). Copenhagen, Denmark, 2020. DOI: 10.4230/LIPIcs.CPM.2020.10. URL: https://hal.inria.fr/hal-03085847.

[34] R. Grossi, A. Marino and S. Moghtasedi. 'Finding Structurally and Temporally Similar Trajectories in Graphs'. In: SEA 2020 - 18th International Symposium on Experimental Algorithms. Vol. 160. Leibniz International Proceedings in Informatics (LIPIcs). Catania, Italy, 2020, pp. 1–13. DOI: 10.4230/LIPIcs.SEA.2020.24. URL: https://hal.inria.fr/hal-02956070.

[35] A. Marchetti-Spaccamela, N. Megow, J. Schlöter, M. Skutella and L. Stougie. 'On the Complexity of Conditional DAG Scheduling in Multiprocessor Systems'. In: IPDPS 2020 - IEEE International Parallel and Distributed Processing Symposium. New Orleans / Virtual, United States, 18th May 2020, pp. 1061–1070. DOI: 10.1109/IPDPS47924.2020.00112. URL: https://hal.inria.fr/hal-03087716.

Conferences without proceedings

[36] G. Bernardini, A. Conte, G. Gourdel, R. Grossi, G. Loukides, N. Pisanti, S. P. Pissis, G. Punzi, L. Stougie and M. Sweering. 'Hide and Mine in Strings: Hardness and Algorithms'. In: International Conference on Data Mining (ICDM). Sorrento, Italy, 17th Nov. 2020. URL: https://hal.archives-ouvertes.fr/hal-03070560.

Doctoral dissertations and habilitation theses

[37] C. Moraga. 'Development of new algorithms to advance on the discovery of microRNAs'. Université Claude Bernard Lyon 1, 3rd Nov. 2020. URL: https://hal.archives-ouvertes.fr/tel-03131632.

[38] I. Ziska. 'Models and algorithms for investigating and exploiting the metabolism of microorganisms'. Université Claude Bernard Lyon 1 (UCBL), 24th Nov. 2020. URL: https://tel.archives-ouvertes.fr/tel-03131655.